Uniciencia Vol. 39(1), pp. 1-21, January-December, 2025
www.revistas.una.ac.cr/uniciencia
revistauniciencia@una.cr
DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Application of the LDA Model for Topic Identification: A Case Study on Dengue Diagnosis

Aplicación del modelo de LDA para la identificación temas: Un caso aplicado sobre diagnóstico de dengue

Aplicação do modelo LDA para identificação de problemas: Um Caso Aplicado sobre Diagnóstico de Dengue

Jennyfer Portilla-Yela[1]*, **Andrés-Felipe Palomino-Montezuma**[2], **Diego-Fernando Manotas-Duque**[1], **José-Rafael Tovar-Cuevas**[2]

Received: Aug/19/2024 • Accepted: Mar/18/2025 • Published: Nov/30/2025

Abstract

[Objective]: This research aimed to identify and analyze topics in the scientific literature related to dengue, with a focus on diagnosis, signs, and symptoms using the Latent Dirichlet Assignment (LDA) model. **[Methodology]:** Articles were collected from various databases, including VHL, Web of Science, Ovid, Scopus, PubMed, Health & Medical, ScienceDirect, and Google Scholar, covering 2000-2024. The search equation was designed using key terms such as "dengue," "signs," "symptoms," and "diagnosis," along with MeSH terms to ensure the inclusion of relevant articles. The LDA model was then implemented to analyze the collected articles. **[Results]:** The implementation of the LDA model identified four main themes: 1) Dengue diagnosis and clinical presentation, 2) Research and control interventions, 3) Severe dengue and its clinical manifestations, and 4) Virus detection, including dengue, zika, and chikungunya. This thematic analysis facilitated the organization and understanding of literature, providing an overview of the predominant themes in dengue research. **[Conclusions]:** The study's approach not only enhanced the organization and understanding of the articles found but also provided insights into the predominant themes in dengue literature, which may guide future research and improve diagnostic and treatment strategies.

Keywords: Text mining; Statistical learning; Text processing; Text recognition; LDA Model; Dengue.

Resumen

[Objetivo]: Esta investigación tuvo como objetivo identificar y analizar temas de la literatura científica relacionados con el dengue, con énfasis en el diagnóstico, signos y síntomas, utilizando el Modelo de Asignación Latente Dirichlet (LDA). [**Metodología]:** Se recopilaron artículos de diversas bases de datos,

* Corresponding author
Jennyfer Portilla-Yela, jennyfer.portilla@correounivalle.edu.co, https://orcid.org/0000-0002-9134-9309
Andrés-Felipe Palomino-Montezuma, palomino.andres@correounivalle.edu.co, https://orcid.org/0009-0000-8577-2197
Diego-Fernando Manotas-Duque, diego.manotas@correounivalle.edu.co, https://orcid.org/0000-0003-0148-9840
José-Rafael Tovar-Cuevas, jose.r.tovar@correounivalle.edu.co, https://orcid.org/0000-0003-0432-4144
1 School of Industrial Engineering, Universidad del Valle, Cali, Colombia.
2 School of Statistics, Universidad del Valle, Cali, Colombia.

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

incluyendo BVS, Web of Science, Ovid, Scopus, PubMed, Health & Medical, ScienceDirect y Google Scholar, cubriendo el período de 2000 a 2024. La ecuación de búsqueda se diseñó empleando términos clave como "dengue", "signos", "síntomas" y "diagnóstico", junto con términos MeSH para asegurar la inclusión de artículos relevantes. A continuación, se implementó el modelo LDA para analizar los artículos recopilados. [**Resultados]:** La implementación del modelo LDA identificó cuatro temas principales: 1) Diagnóstico y presentación clínica del dengue, 2) Intervenciones de investigación y control, 3) Dengue grave y sus manifestaciones clínicas, y 4) Detección del virus, incluyendo dengue, zika y chikungunya. Este análisis temático facilitó la organización y comprensión de la literatura, proporcionando una visión general de los temas predominantes en la investigación del dengue. [**Conclusiones]:** El enfoque del estudio no solo mejoró la organización y comprensión de los artículos encontrados, sino que, también proporcionó una visión de los temas predominantes en la literatura sobre el dengue, lo que puede orientar a las futuras investigaciones y mejorar las estrategias de diagnóstico y tratamiento.

Palabras claves: Aprendizaje estadístico; Dengue; Minería de textos; Modelo LDA; Procesamiento de textos; Reconocimiento de textos.

Resumo

[Objetivo]: Esta pesquisa teve como objetivo identificar e analisar temas da literatura científica relacionadas à dengue, com ênfase no diagnóstico, nos sinais e nos sintomas, utilizando o Modelo de Alocação de Dirichlet Latente (LDA). **[Metodologia]**: Os artigos foram coletados de vários bancos de dados, incluindo BVS, Web of Science, Ovid, Scopus, PubMed, Health & Medical, ScienceDirect e Google Scholar, abrangendo o período de 2000 a 2024. A equação de pesquisa foi elaborada usando termos-chave como "dengue", "sinais", "sintomas" e "diagnóstico", juntamente com termos MeSH para garantir a inclusão de artigos relevantes. O modelo LDA foi então implementado para analisar os artigos coletados. **[Resultados]:** A implementação do modelo LDA identificou quatro temas principais: 1) Diagnóstico e apresentação clínica da dengue; 2) Intervenções de pesquisa e controle; 3) Dengue grave e suas manifestações clínicas; e 4) Detecção de vírus, incluindo dengue, Zika e Chikungunya. Essa análise temática facilitou a organização e a compreensão da literatura, fornecendo uma visão geral dos temas predominantes na pesquisa sobre a dengue. **[Conclusões]:** A abordagem do estudo não apenas melhorou a organização e a compreensão dos artigos encontrados, mas também forneceu uma visão dos temas predominantes na literatura sobre dengue, o que pode orientar pesquisas futuras e melhorar as estratégias de diagnóstico e tratamento.

Palavras-chave: Mineração de texto; aprendizado estatístico; processamento de texto; reconhecimento de texto; modelo LDA; dengue.

Introduction

According to Luquea *et al*. (2021), technological advances and advanced search engines facilitate the rapid acquisition of information on a specific topic. However, access to scientific content in online databases, although diverse, presents difficulties in synthesizing trends in scientific research in a particular area (Gulo & Rúbio, 2015; Trueba-Gómez & Estrada-Lorenzo, 2010). According to Asmussen and Møller (2019), exploratory manual reviews are characterized as a laborious process that requires considerable time and is limited by the available processing capacity, leading to a

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

reduced analysis of articles. While there are several ways to conduct an exploratory review, most methods require a considerable initial investment of time and pre-existing knowledge (Asmussen & Møller, 2019).

For this reason, literature proposes techniques, such as those from the field of text mining, which enable exploring the vast number of available articles in a more efficient and structured manner. One of these techniques is the Latent Dirichlet Allocation Model (LDA), which represents a sophisticated information retrieval tool. This model automatically identifies general themes in a set of text documents and attempts to uncover implicit themes, thereby facilitating the automatic organization, understanding, searching, and summarizing of numerous documents.

In Asmussen and Møller (2019), the application of the LDA model is suggested as a transparent, reliable, fast, and reproducible process for reviewing large numbers of documents, an essential task in constructing state-of-the-art research papers. This article presents the LDA model as a valuable tool that provides an overview of topics during the exploration and literature review phases, minimizing the effort required before tackling time-consuming manual reviews. This approach is particularly beneficial for novice researchers or those with limited knowledge in a specific research field.

The LDA model has previously been used to identify concepts and topics as mentioned in the review presented by Asmussen and Møller (2019), which indicates that most cases have been employed to analyze web content (Ghosh & Guha, 2013; Guo *et al*., 2016; Elgesem *et al*., 2016, 2019; Parra *et al*., 2016), newspaper articles (Koltsova & Koltcov, 2013; DiMaggio *et al*., 2013; Van Atteveldt *et al*., 2014; Evans, 2014; Jacobi *et al*., 2018), books (Jockers & Mimno, 2013), speeches (Quinn *et al*., 2010) and, in one case, videos (Baum, 2012). The application of the LDA model in medical research is demonstrated in several studies. For example, Sahria and Fudholi (2020) employed it to analyze themes in Indonesian health research. The authors applied the model to analyze news headlines over eight months of the COVID-19 pandemic; Luquea *et al*. (2021) used a similar approach to examine patterns in COVID-19 scientific research from abstracts published in PubMed during the first half of 2020. Another example is the study by Rojas (2022) on cervical cancer in social networks, which used the LDA model to identify issues such as the effectiveness of HPV vaccines, the relationship with other diseases, and medical programs. Finally, Guzman-Ponce *et al*. (2023) proposed its use to investigate COVID-19 survival factors in Mexico, using data from the Ministry of Health.

Similarly, topic modeling has been applied to study vector-borne diseases such as dengue. In Arenas Silva's (2022) work, tweets about dengue in Colombia between January 2019 and January 2021 were collected using the Twitter API and Python. The results showed that most user discussions focused on prevention, control, and mitigation initiatives. In addition, the topic analysis identified that March was the month with the highest activity in posts about dengue, followed by January and February, which coincides with the trend of reported cases in the same period.

Although the LDA model has demonstrated its usefulness in various areas, its application in dengue studies is especially relevant due to the significant impact of this disease on global public health. Dengue can affect people of all ages and present a broad spectrum of manifestations, ranging from

UNICIENCIA Vol. 39, N° 1, pp. 1-21, January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

asymptomatic infections to severe and potentially fatal forms (Guzman *et al*., 2010). It is currently endemic in more than 100 countries, mainly in tropical and subtropical regions such as South America, Central America and the South Pacific, making it one of the fastest-spreading vector-borne diseases.

According to a World Health Organization (WHO) (2023) report, the global incidence of dengue has increased dramatically over the past two decades, representing a critical public health challenge. Between 2000 and 2019, the number of reported cases worldwide increased tenfold, from 500,000 to 5.2 million. Given its rapid spread and increasing impact on public health, the study of dengue is essential to optimize diagnosis, strengthen epidemiological surveillance, and develop more effective control strategies.

The objective of the article is to employ the LDA model to effectively identify and analyze the topics present in the literature related to dengue diagnosis, due to the volume of articles that can be presented. This approach will not only facilitate the segmentation of available information but also enhance orientation to identify and address knowledge gaps in the field of dengue research. By clearly understanding the predominant themes in this research, the synthesis of information will be facilitated, and underexplored areas may be identified, suggesting opportunities for future research.

The results obtained from the LDA model identified four main themes within the scientific literature on dengue: diagnosis and clinical presentation, investigation and control interventions, severe dengue and its clinical manifestations, and virus detection (including dengue, zika, and chikungunya). Each of these topics was found to encompass key aspects in the study of the disease, ranging from diagnostic methods and clinical symptoms to control strategies and epidemiological surveillance.

The paper is structured as follows: section 2 presents the theoretical framework and rationale of the LDA model, as well as its applicability in scientific studies; section 3 describes the methodology employed, including the process of data collection, text preprocessing and configuration of the LDA model; section 4 presents the results obtained and their interpretation in the context of dengue bibliometric analysis; section 5 discusses the most relevant findings and their possible implications for dengue research; finally, section 6 presents a summary of the contributions of the study and offers recommendations for future research.

Theoretical Framework

Model LDA

The Latent Dirichlet Allocation (LDA) model is an unsupervised learning model, meaning there is no prior information about the possible topics or themes, or at least they are not predefined. The LDA model is based on a latent variable model derived from the interaction between the observed data and hidden random variables. In this context, the observed data correspond to each text, while the latent variables represent the topics associated with each document (Silvestre, 2018).

Each document, or bag of words, is considered a mixture of several topics assigned by the same algorithm; in this part, it is assumed that the distribution of each topic comes from a Dirichlet probability distribution. This means that the algorithm allows each bag of words (Document) to belong simultaneously to different topics, with a different weight for each topic; that is,

UNICIENCIA Vol. 39, N° 1, pp. 1-21, January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

each bag of words is more likely to belong to a particular topic than to another (Diaz Rubiano, 2022).

According to Diaz Rubiano (2022), the LDA model has its probabilistic basis in Bayesian statistics; this model makes better use of the available data without the need to collect large amounts of information, thanks to its Bayesian foundation (Blei *et al.*, 2003). The LDA model is based on some basic assumptions. First, it assumes that documents dealing with similar topics use similar words in their content and that topics are distributions of a fixed vocabulary, K topics. These basic assumptions underscore the ability of the LDA model to discover hidden patterns in document sets and help identify underlying topic structures.

It is important to define the following terms before proceeding (Blei *et al.*, 2003):

- Corpus: They represent the set of texts or documents in which we are interested in discovering topics. In our case, it is the set of article summaries. It is denoted by: D = $\{d_1, d_2, \ldots, d_n\}$, each dj represents the *j-th* document.
- Topics: These are latent topics or underlying categories that can describe the content of the documents. The LDA model assumes that each document is a mixture of several topics. It is denoted by: $Z = \{z_1, z_2, \ldots, z_k\}$, where k represents the number of topics.
- Words: They are the smallest units in documents. The LDA models the generation of words in a document from its topics. It is denoted by w_i, where the subscript i indicates the *i-th* word of the finite collection of words called vocabulary $w = \{w_1, w_2, \ldots, w_N\}$.
- Term-Document Matrix: It represents the relationship between terms (words) and documents. Each row of the matrix corresponds to a term, each column to a document, and each cell contains the frequency or weight of the term in the corresponding document. This matrix is fundamental to analyzing and discovering the structure of topics.

Considering a D corpus including a total of M documents, where each j document with $j = 1, 2, \ldots, M$ consists of N_j words, the Latent Dirichlet Allocation (LDA) model addresses the modeling of this D corpus. Figure 1 shows how the LDA model works.

Methodology

Article procurement

Documents will be searched in BVS, Web of Science, Ovid, Scopus, PubMed,

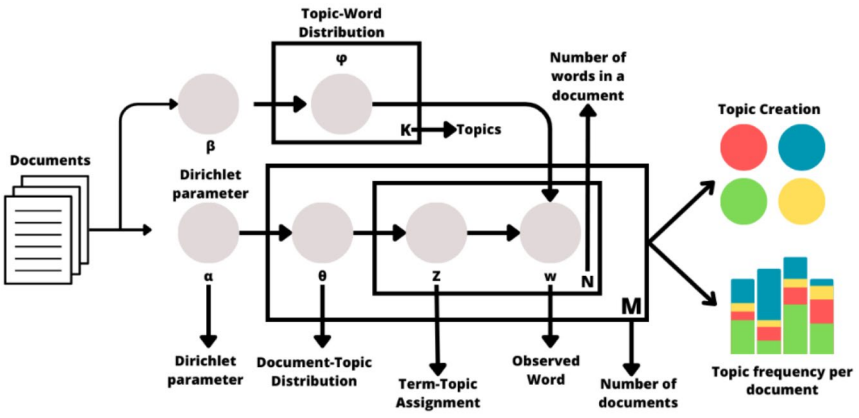


Figure 1. *How does an LDA model work?*
Source: Arteaga and Mendoza (2023).

UNICIENCIA Vol. 39, N° 1, pp. 1-21, January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Health & Medical, ScienceDirect, and Google Scholar databases. The search strategy will be based on the equation used by the PubMed database, which offers a vast collection of peer-reviewed medical articles and advanced filters. The equation will be modified to include or exclude Mesh terms as required by different search engines. For the construction of the search equation, an expert with extensive experience in public health research and epidemiology, who has worked on projects related to dengue, environmental health, and diagnostic tests, was consulted. Her experience was key to verifying the inclusion of relevant terms in the study. The search equation is as follows:

((Dengue) OR ('Dengue' [Mesh]) OR ('Breakbone Fever')) AND ((Diagnosis) OR ('Diagnosis' [Mesh]) OR ('Diagnoses and Examinations')) AND (('Signs and Symptoms') OR ('Symptom Assessment'[Mesh]))

Initially, 1,677 documents were obtained from all the online databases consulted, covering the period from 1855 to 2024. It was decided to use documents from 2000 to 2024 to cover a broad and up-to-date period. The selection criterion for selecting documents was to include all those of type Article and exclude all "Literature reviews". This was done through Rayyan (A collaborative web application developed by Qatar Computing Research Institute to streamline literature reviews). Finally, 1,224 documents were obtained in this initial phase, as the article's central theme is dengue. Articles containing the word 'dengue' in the title or abstract were selected, and a total of 938 articles were obtained, characterized by their title and, subsequently, their abstract, which was modeled using an LDA model.

Article preparation

For the analysis of frequencies in titles and abstracts presented in sections 3.1.1 and 3.1.2, steps 1 to 7 were used to remove them (Figure 2). Similarly, the abtracts for the construction of the LDA model of Section 6 were removed.

The steps applied in this procedure are described below (see Figure 2):

1. **Conversion of the text to lowercase**: To avoid inconsistencies due to differences in the use of uppercase and lowercase letters, the entire text was converted to lowercase. This ensures that words such as 'DENGUE' and 'dengue', for example, are treated as the same entity, avoiding unnecessary duplications in the analysis.

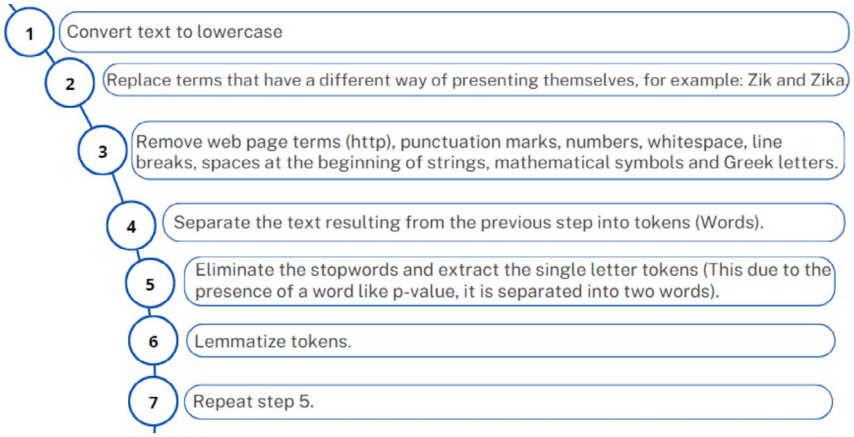


Figure 2. *Stages of the preparation process.*
Source: own research.

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

2. **Standardization of terms with different spellings:** Some terms may be presented differently within the dataset, either through abbreviations, spelling variations, or variations in the health literature for referring to them. To unify these representations, terms with different spellings that refer to the same entity, such as 'zik' and 'zika', were replaced, ensuring consistency in topic modeling.

3. **Elimination of unnecessary elements:** All characters and elements that do not provide relevant information for the analysis of topics were eliminated, including:

   - Web page terms (such as http or links embedded in the texts).
   - Punctuation marks, since they do not contribute semantic meaning in topic modeling.
   - Numbers and mathematical symbols, because these can generate noise in the detection of relevant topics.
   - Blank spaces and line breaks in texts, avoiding segmentation problems in the analysis.
   - Greek letters and special characters, which are not representative in the interpretation of topics.

4. **Tokenization:** This is a process in which the text is divided into individual units called tokens (words). Each token represents a word within the document and serves as the basis for constructing the LDA model.

5. **Elimination of stop words and single-letter tokens:** Stop words, which are terms with high frequency but low semantic value in the analysis, such as the, the, of, and in, were eliminated. The list of stop words in the R stop words package, as compiled by Benoit et al. (2021), was utilized, and additional irrelevant terms that recur in scientific abstracts were included, such as 'abstract,' 'background,' "conclusion," 'copyright,' 'discussion,' 'methodology,' and others commonly found in article abstracts. Single-letter tokens were also eliminated, which is important, because words such as 'p-value' can be split into two tokens (p and value), which would lead to errors in identifying significant terms.

6. **Lemmatization of tokens:** Lemmatization consists of reducing words to their base or root form. A more precise linguistic analysis is used, returning the correct grammatical form of each word. This improves the consistency of the analysis by grouping terms with the same meaning.

7. **Repeated removal of empty words:** After lemmatization, some empty words that may have changed form reappear in the texts. To make sure that these do not affect the topic modeling, a second stopword cleaning was performed.

Selection of the number of topics

To determine the optimal number of themes in the LDA model, a combination of quantitative techniques and visualization was applied to ensure that the extracted themes were interpretable and consistent.

First, perplexity was calculated over a range of 2 to 20 subjects (Graph 7). According to Jiang (2023), a lower perplexity

UNICIENCIA Vol. 39, N° 1, pp. 1-21, January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

indicates that the model assigns a higher probability to the data set, resulting in better predictive ability. Therefore, the elbow method was used to identify the point at which the reduction in perplexity stabilizes, achieving a balance between granularity and interpretability of the subjects. In addition, the metrics implemented by Nikita (2020) are used, which correspond to those in Arun et al. (2010), Cao et al. (2009), Deveaud et al. (2014), and Griffiths and Steyvers (2004).

To further validate this selection, biplots were generated for different configurations of themes (The selected biplot is in Graph 9). This visual representation allowed examining how the themes are distributed in a two-dimensional space. The biplot was produced using the LDAvis package (Sievert & Shirley, 2015). According to Pilacuan-Bonete *et al*. (2022), LDAvis enables the visualization of themes generated by the LDA model, representing them by circles. The diameter of each circle is proportional to the number of words assigned to that topic within the model, i.e., the larger the diameter, the greater the proportion of words associated with that topic.

These circles are arranged in a multidimensional scaling (MDS) plane, where the distribution reflects the map of distances between topics in a two-dimensional space. The position of the circles is determined by a multidimensional scaling algorithm, based on the similarity of words between topics; thus, the closest topics in the graph share a greater number of terms (Pilacuan-Bonete *et al*., 2022).

Resulted

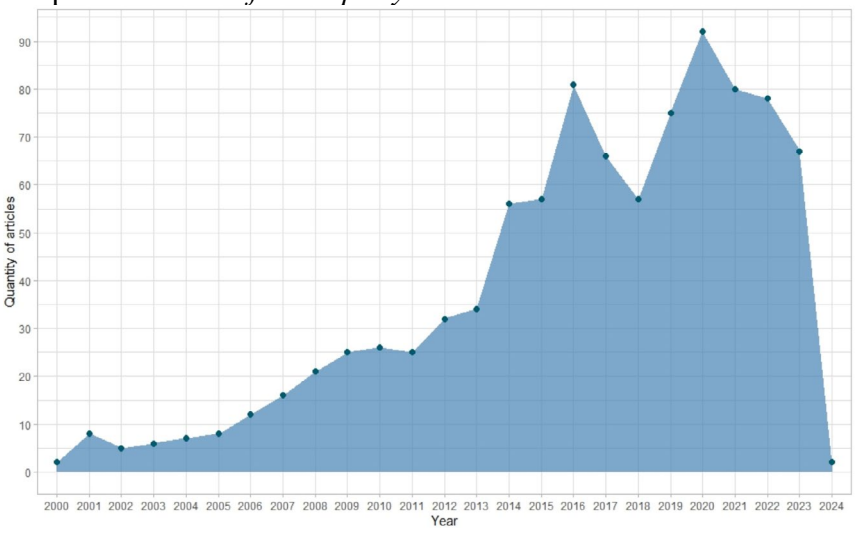
Characterization of articles and journals

Graph 1 illustrates the production of articles between 2000 and 2024 (January). At the beginning of the period, production is low, with only two articles, but increases progressively until 2013, where a considerable increase is observed, reaching a first peak in 2016 (with 81 articles) and a second peak in 2020 (with 92 articles).

This increase in scientific output coincides with growing global concern about dengue, as reported by the World Health Organization (WHO) in its 2023 report. In 2019, an unprecedented peak was recorded, with cases reported in 129 countries. The increase in the number of infections and the spread of the disease to new regions prompted research, accompanied by a corresponding increase in focus on surveillance, diagnosis, and control strategies.

According to the same 2023 report, although there was a temporary decline in cases between 2020 and

Graph 1. *Number of items per year*.



Source: own research.

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

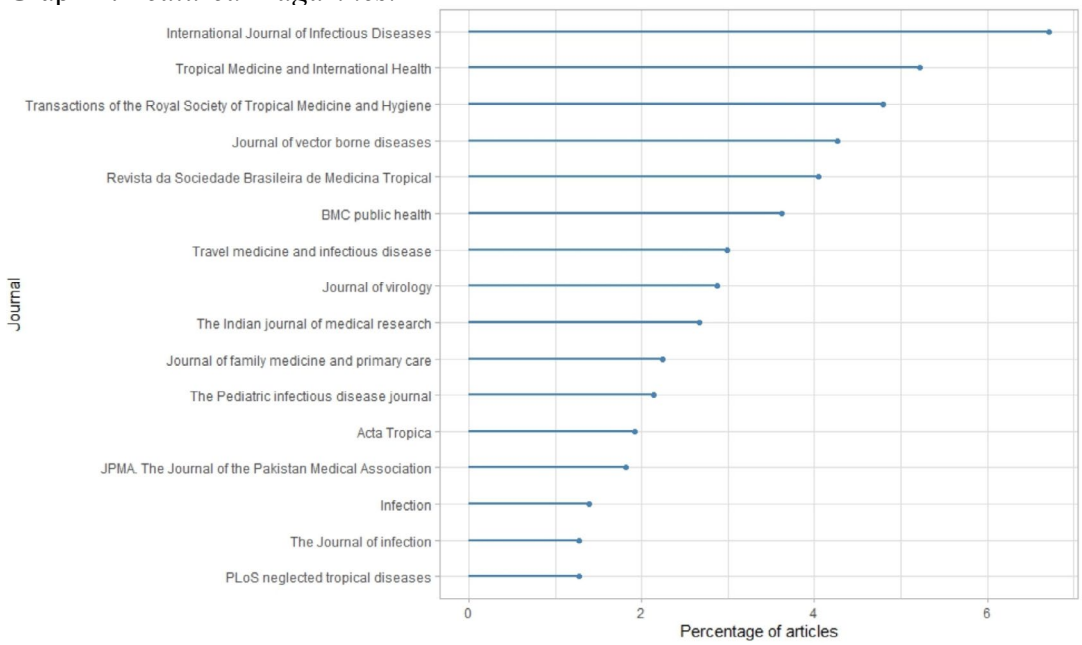
2022, attributed to the COVID-19 pandemic and a lower reporting rate, a global resurgence of dengue was observed in 2023. This resurgence was characterized by a significant increase in the number and magnitude of cases, as well as simultaneous outbreaks in previously unaffected regions. These epidemiological patterns could explain the variations in scientific output, reflecting increased research interest in response to major disease outbreaks.

Graph 2 shows the top 15 journals with the highest percentage of articles. It is observed that the journal International Journal of Infectious Diseases has, for the selected articles, the highest SCImago journal rank (SJR) indicator, which suggests it has a more significant influence and importance compared to other journals in its area (Medicine) that are indexed in Scopus. Another journal that stands out is the Journal of Virology, which has an SJR impact of 1,795, the highest H-index (315), and ranks third in total citations over the last three years. This suggests that these journals have published many high-quality, widely cited articles, contributing to their outstanding reputation.

Analysis of word frequency in titles

When performing a frequency analysis of the words present in the titles, it was found that the word "dengue" had a frequency of 719 times, a situation that was expected, given that the presence of the word in the title was an inclusion criterion for the final articles. Based on the words with the highest frequency (Graph 3), it can be mentioned that the selected articles seem to explore aspects such as the epidemiology of dengue and other diseases like zika or chikungunya, which suggests a consideration of the possible interactions of mosquito-borne diseases, additionally, their presence in different regions such as India, Brazil or endemic, as well as their impact on

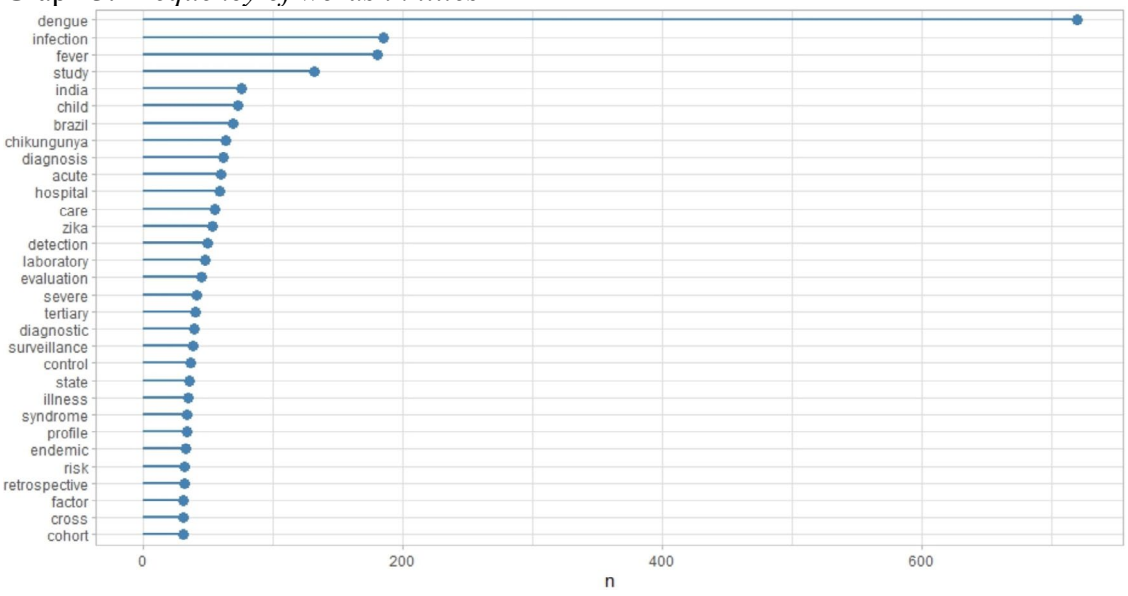
Graph 2. *Featured Magazines.*



Source: own research.

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Graph 3. *Frequency of words in titles*

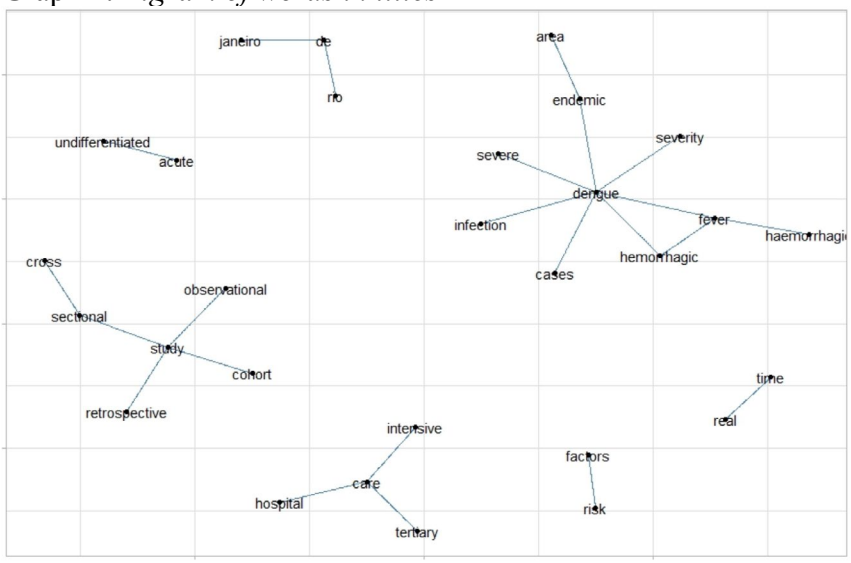


Source: own research.

children. Finding words such as diagnosis, fever, hospital, care, detection, risk, and laboratory suggests that we may be looking at articles covering research related to medical care in hospital settings and others dedicated to the detection, diagnosis, and evaluation of the disease through laboratories.

Graph 4 shows a bigram for the words included in the titles, indicating that there are groups of related words but not a connection among all of them; this could be interpreted as the presence of a diversity of topics within the context we are dealing with. There is a notable co-occurrence among the world's 'study', 'cross', 'sectional', 'cohort', 'observational', and 'retrospective', connected to the word study. This pattern indicates the inclusion of diverse methodological approaches in the reviewed articles, as evidenced by the various study types mentioned, including cross-sectional, cohort, observational, and retrospective studies.

Graph 4. *Bigram of words in titles*



Source: own research.

UNICIENCIA Vol. 39, N° 1, pp. 1-21, January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

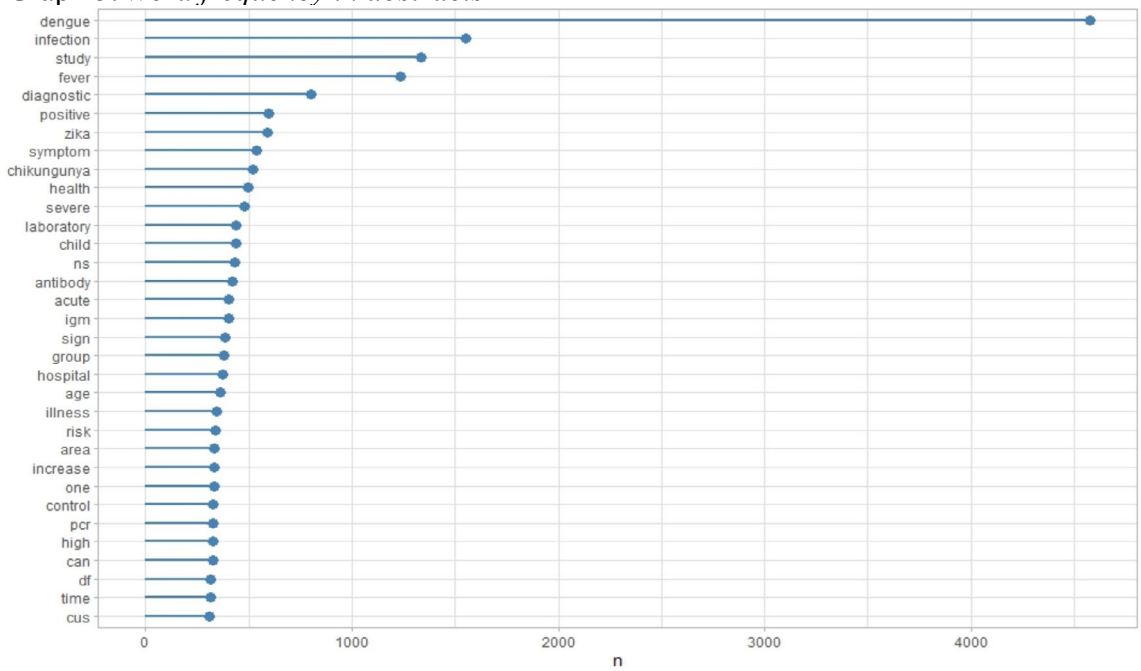
DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Another segment highlights the relationship between the words ‘care’, ‘hospital’, ‘intensive’, and ‘tertiary’, all of which are connected by the word ‘care’, suggesting that they may encompass articles about medical care and attention. The next segment displays the words ‘undifferentiated’ and ‘acute’, which may refer to an undifferentiated acute syndrome or disease. Additionally, co-occurrences are observed between the words ‘risk’ and ‘factors’, as noted by the Ministerio de Salud y Protección Social – Federación Médica Colombiana (2013). Dengue is an acute viral disease that has various clinical forms, from asymptomatic undifferentiated presentations to severe forms leading to shock and vital organ failure. Finally, the co-occurrence of the word’s ‘dengue’, ‘area’, ‘infection’, ‘endemic’, ‘severity”, ‘fever’, and ‘hemorrhagic’ suggests that the titles of the articles have a thematic focus around dengue, a situation to be expected due to the search equation and inclusion criteria of the selected articles. These articles cover geographical aspects (due to the word ‘endemic’), dengue itself, the severity of the disease, symptoms such as fever, and more severe forms, including dengue hemorrhagic fever.

Word frequency analysis in abstracts

Graph 5 shows the frequency of word occurrence in the abstracts, again highlighting the words ‘dengue’, ‘infection’, ‘study’, ‘fever’, and ‘diagnostic’, among others. If we compare it with the distribution of the words in the title presented in Graph 3, we find a similar composition. One interpretation of this set of words is that the abstracts of the articles tend to focus on research related to mosquito-borne diseases, particularly dengue, addressing topics such as diagnosis, laboratory techniques, risk factors, and the prevalence of the disease in children.

Graph 5. *Word frequency in abstracts*



Source: own research.

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

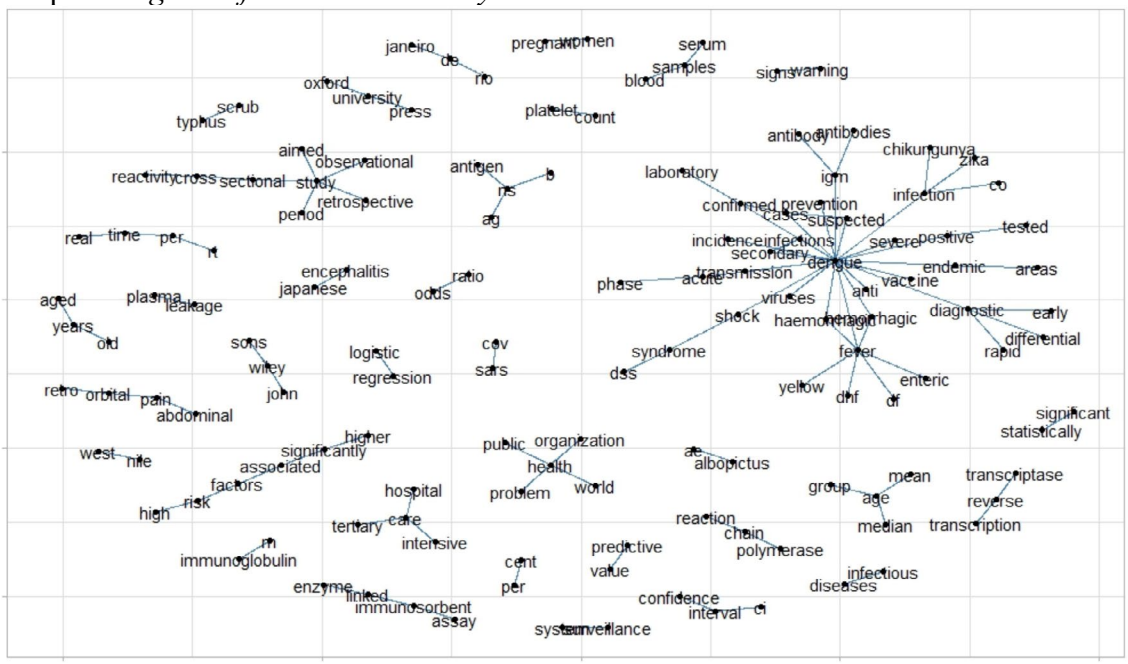
According to the bigram of Graph 6, the co-occurrence of the word 'dengue' with other words referring to diseases such as zika and chikungunya, and the word 'co-infection', stands out, which suggests that some articles may be referring to cases or situations where both diseases are present. The relationship with the words 'diagnosis,' 'early,' 'rapid,' and 'differential' is also evident. Topics related to platelet count, plasma, and other medical terms, such as blood samples, antigens, ns, and alarm signs, are identified. Additionally, words associated with statistical topics such as 'logistic regression', 'confidence interval', 'percentage', 'mean', 'median', 'statistical significance', and 'predictive value' were identified.

LDA model for abstracts

Initially, we proceeded to clean up the article summaries, eliminating punctuation marks, mathematical symbols, empty words, or stop words from the package stop words Benoit *et al.* (2021), and additional words were included: abstract, background, conclusion, copyright, discussion, and methodology, among others, which were very frequent in article abstracts. According to the results of Graph 7, it is obtained that the optimal number could be three because it is the first inflection point of the graph; now, based on the results of Graph 8, the number of topics could be from 10 to 14; however, it is observed that the number of topics where the Griffiths (2004) and Cao Juan (2009) metrics jointly become minimal is three, while for Arun (2010) and Deveaud (2014) metrics it is 4, 8, 10.

The LDA model is fitted considering different numbers of topics. A biplot is generated for each topic configuration, and it is visually verified that the formed topics are mutually exclusive. This exclusivity is especially evident when comparing the cases of groups topics 3 and 4. After reviewing

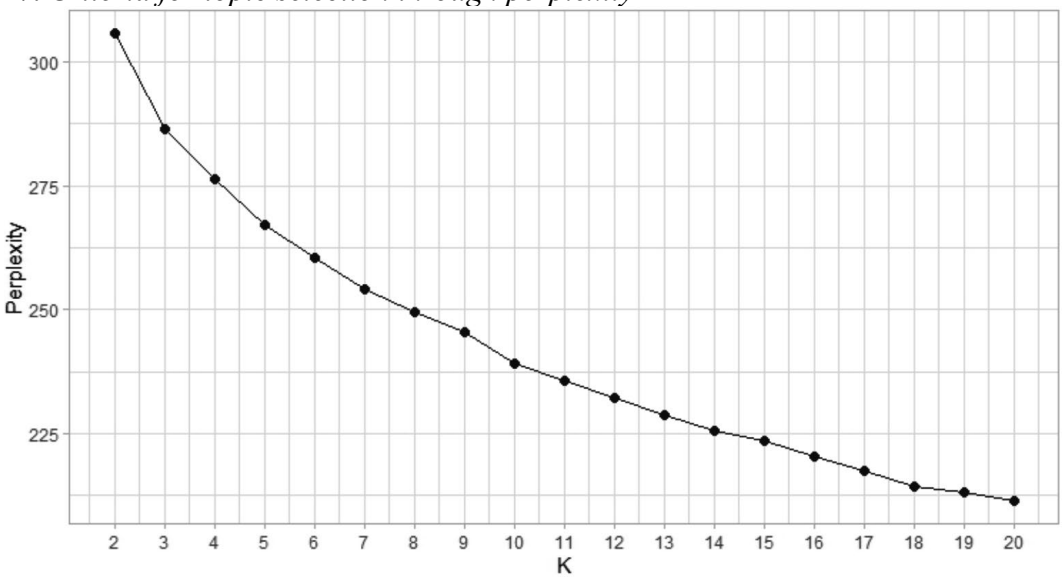
Graph 6. *Bigram of words in summary*



Source: own research.

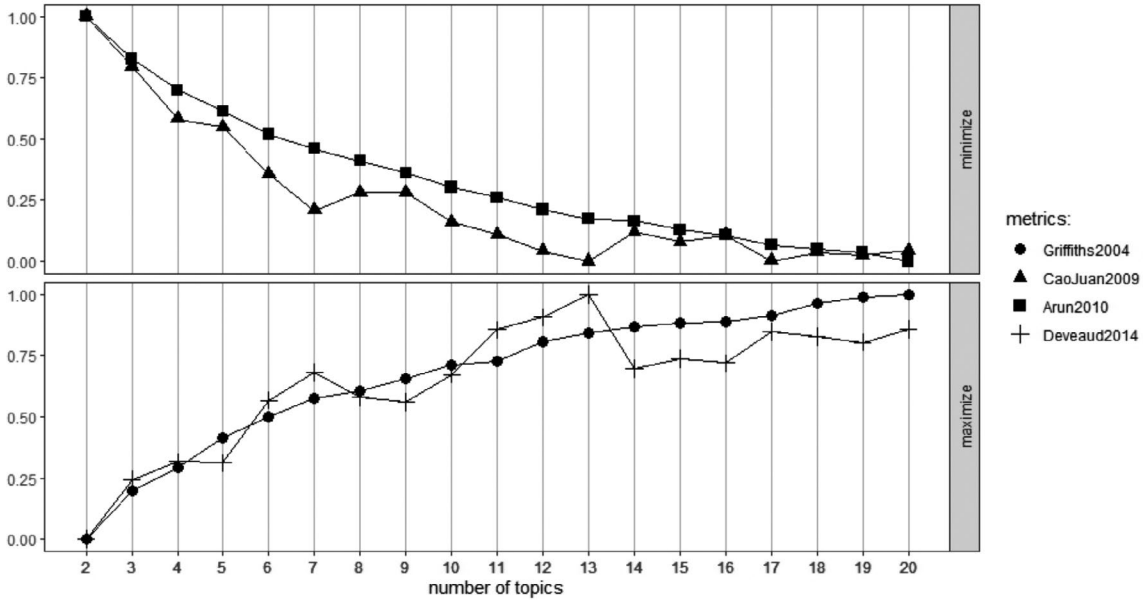
DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Graph 7. *Criteria for topic selection through perplexity*



Source: own research.

Graph 8. *Criteria for topic selection using other metrics*



Source: own research.

UNICIENCIA Vol. 39, N°. 1, pp. 1-21. January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

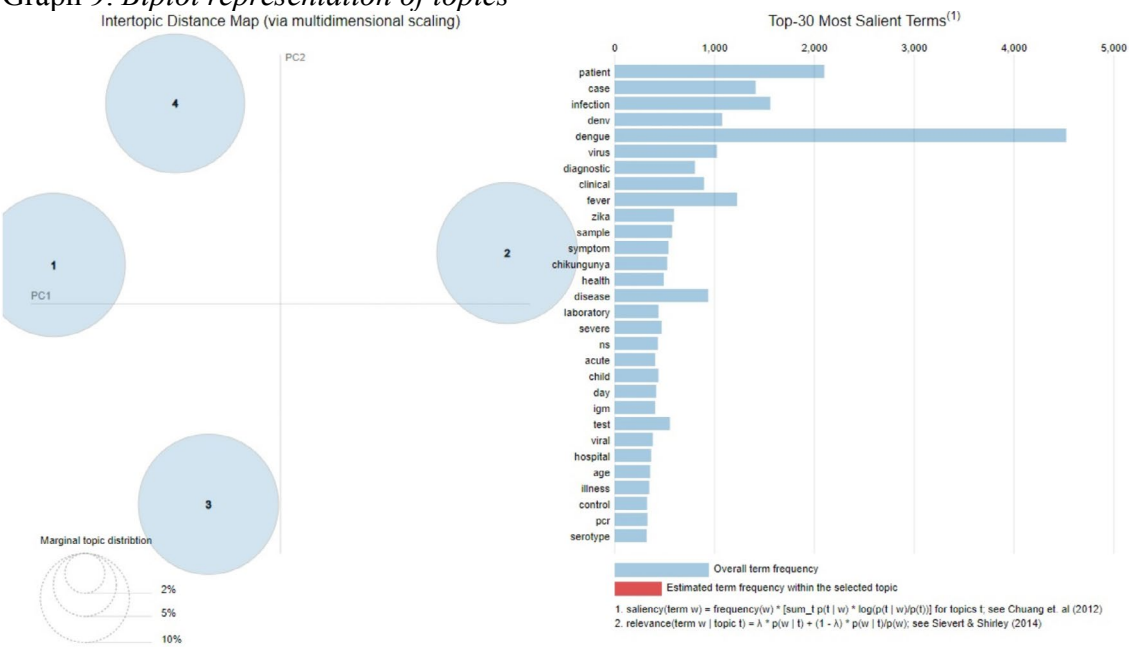
the representations graphically, the decision is made to select the optimal number of topics, four. In Graph 9, the topics conformed with *k=4* are mutually exclusive. In Graph 9, the topics are represented by circles in a two-dimensional plane; their size reflects the importance of the topic in the set of articles (Abstracts); in this case, there is no noticeable difference between the topics.

Finally, the proposed LDA model allocated 22.42%, 23.80%, 27.95%, and 25.50% of articles to each topic. Graph 10 shows the top 30 words best presented in each topic, with which we will proceed to name them in the following sections.

Topic 1: Diagnosis and clinical presentation

The keywords 'case', 'dengue', 'diagnostic', 'clinical', 'fever', 'symptom', 'disease', and 'test' suggest that this topic primarily focuses on the diagnosis and clinical presentation, with a greater emphasis on 'dengue'. This idea is reinforced by words such as 'laboratory', 'blood', 'positive', 'negative', 'study', indicating a connection with laboratory tests, and 'results'. Furthermore, using terms such as 'child', 'acute', and 'year' implies that the subject may pertain to the variation in Dengue presentation across different age groups, including children, acute syndromes, and various time frames. To explore this topic, a review of the five most relevant articles was conducted in order of significance: Marks *et al.* (2016), Kalra *et al.* (2016), Laoprasopwattana *et al.* (2020), Herbinger *et al.* (2011), and da Silva Ferreira *et al.* (2020). The five main articles discuss disease studies, such as typhoid fever, dengue, and chikungunya. They employ research methods such as retrospective or prospective studies, data analysis, and pilot studies. Each article identifies and differentiates diseases through clinical and laboratory diagnostic methods.

Graph 9. *Biplot representation of topics*

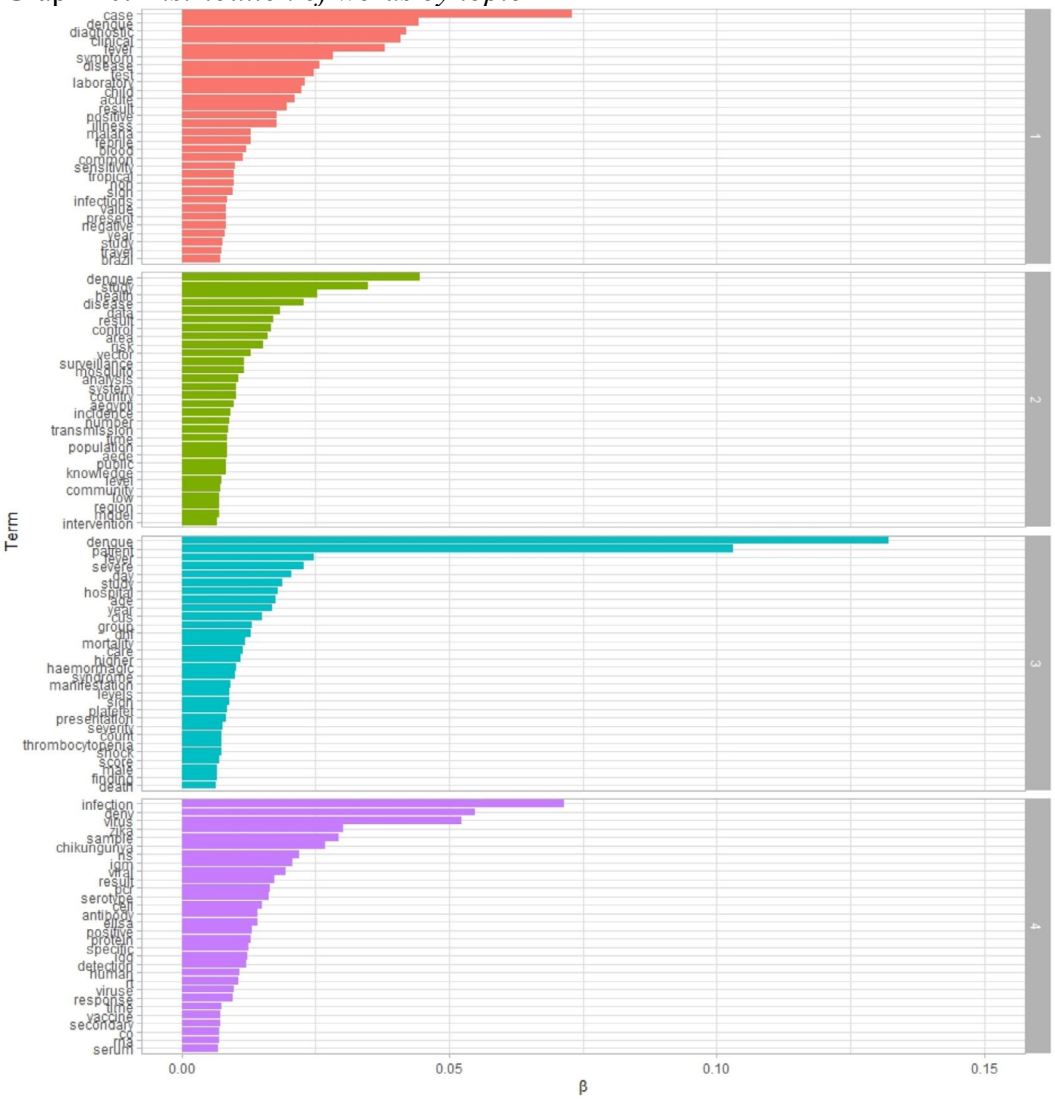


Source: own research.

UNICIENCIA Vol. 39, N° 1, pp. 1-21, January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistauniciencia@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Graph 10. *Distribution of words by topic*



Source: own research.

Topic 2: Research and Control Interventions

Keywords such as ‘dengue’, ‘study’, ‘health’, ‘disease’, ‘data’, ‘result’, ‘incidence’, ‘transmission’, ‘knowledge’, and ‘model’ suggest that this topic focuses on dengue research. In contrast, words such as ‘control’, ‘area’, ‘risk’, ‘vector’, ‘surveillance’, ‘mosquito’, ‘population’, ‘community’, and ‘intervention’ indicate a connection with disease control strategies, geographical risk areas, vectors (such as the Aedes aegypti mosquito), and surveillance systems. A review of the five most representative articles on the topic is performed in order of representativeness: Sanchez *et al*. (2005), Nazareth *et al*. (2014), Jones *et al*. (2014), Schultes *et al*. (2021), and Nivedita (2016). The five main articles have in common that they deal with studies to understand, evaluate, or improve dengue control strategies in different regions such as Cuba, Madeira, and Belo Horizonte (Brazil). Each

UNICIENCIA Vol. 39, N°. 1, pp. 1-21. January-December, 2025 • www.revistas.una.ac.cr/uniciencia • revistaunicienciа@una.cr

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

article presents specific interventions designed to control the spread of the dengue virus, either through policy changes (Sanchez *et al*., 2005), evaluation of community perceptions (Nazareth *et al*., 2014) and (Nivedita, 2016), implementation of new tools such as insecticidal screens (Jones *et al*., 2014), and spatial pattern analysis or health education programs (Schultes *et al*., 2021).

Topic 3: Severe Dengue and Clinical Manifestations (Signs and Symptoms)

Keywords such as 'dengue', 'patient', 'fever', 'severe', 'severity', 'dhf' (dengue hemorrhagic fever), 'haemorrhagic', 'syndrome', 'day', 'study', 'death', 'manifestation', 'care', 'shock', 'platelet', and 'thrombocytopenia' suggest that this topic focuses on aspects related to severe cases of dengue, the clinical manifestation of the disease and possible complications. The presence of terms such as 'hospital', 'age', and 'group' suggests a connection with the presence of different age groups, genders, and the hospital environment. A review of the five most representative articles on the topic was conducted in order of representativeness: Parkash *et al*. (2010), Iqtadar *et al*. (2017), Khetpal *et al*. (2021), Thomas *et al*. (2010), and Khurram *et al*. (2016). The top five articles focus on research dealing with clinical manifestations of dengue, such as plasma leakage [29], progression of dengue to severe cases (Khetpal *et al*., 2021), (Thomas *et al*., 2010), and hepatic manifestations, which can be seen in Parkash *et al*. (2010), and Iqtadar *et al*. (2017).

Topic 4: Virus Detection (Dengue, Zika or Chikungunya)

Keywords such as 'infection', 'denv' (dengue virus), 'virus', 'zika', 'sample', and 'chikungunya' suggest that this topic focuses on studies on diseases transmitted by mosquitoes that could be targeted with diagnostic tests due to the presence of words like 'ns', 'igm', 'igg', 'pcr', 'elisa', 'protein', 'detection', 'cell', ' human', 'positive', and 'serum' in the same way connected with the characterization of the virus by the presence of words like 'serotype' and 'rna'. Now, including words like 'response', 'time', and 'vaccine' could indicate the inclusion of research on vaccination strategies. Finally, the word co could indicate the presence of research that mentions co-infection. The five most representative articles on the topic were reviewed in the order of representativeness; they are Chatel-Chaix *et al*. (2015), Morizono and Chen (2014), Xie *et al*. (2015), Chiang *et al*. (2015), and Pattabhi *et al*. (2016).

The five main articles deal with Dengue viruses, but others mention zika and chikungunya. The articles are focused on understanding and addressing viral infections, whether by mapping critical determinants for replication (Chatel-Chaix *et al*., 2015), exploring molecules related to viral infection (Morizono and Chen, 2014; Xie *et al*., 2015; Pattabhi *et al*., 2016), or designing RNA agonists with enhanced inflammatory antiviral properties against influenza, dengue, and chikungunya viruses (Chiang *et al*., 2015).

Conclusions

During this research, an analysis of dengue-related articles was conducted, focusing on the characterization and modeling of their abstracts. The development of a search equation with MeSH terms enabled access to various health and medical research databases, including sources such as ScienceDirect, which covers a broad spectrum of scientific disciplines. From this review, four

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

main categories of articles were identified: diagnosis and clinical presentation, research and control interventions, severe dengue and clinical manifestations, and virus detection (dengue, zika, or chikungunya).

Although the use of the LDA model has proven to be a valuable tool for synthesizing and organizing the literature on dengue, the present investigation has certain limitations that should be considered. First, the analysis is based on article abstracts only, which implies that the modeling did not include some relevant details on methodologies, specific findings, and discussions. This may affect the accuracy of the thematic classification and limit the scope of the analysis. To address this limitation, future research could expand the dataset by incorporating the full text of the articles. While this would represent a computational and methodological challenge, it would allow for a deeper understanding of each identified topic and facilitate the analysis of key contextual information not usually present in the abstracts.

Furthermore, the LDA model, although widely used in literature, has inherent limitations, such as the need to specify the number of topics a priori and its dependence on the quality of data preprocessing. Alternatively, future research could integrate other analytical methods, such as Latent Semantic Analysis (LSA) or Probabilistic Latent Semantic Analysis (pLSA), which offer different approaches to topic identification and could provide complementary insights to the LDA model. Also, the application of deep learning techniques, such as Topic Modeling based on neural networks (e.g., BERTopic or Transformer-based models), could improve the detection of more complex semantic relationships between terms and enrich the analysis of the underlying topics in the dengue literature.

Finally, the results obtained underline the importance of using unsupervised learning tools, such as the Latent Dirichlet Allocation (LDA) model, in literature reviews. This methodology has proven instrumental in managing large volumes of scientific papers, optimizing the review process, and allowing researchers to focus on the studies most relevant to their lines of research. In this case, we started with a set of 938 articles, which were efficiently segmented into four groups of 211, 224, 263, and 240 documents, respectively. This segmentation facilitates a more detailed review of each category of study, providing a structured basis for future research in the field of dengue and other vector-borne diseases.

Acknowledgment

We thank Dr. Diana María Caicedo, assistant professor and researcher at the Department of Public Health and Epidemiology of the Pontificia Universidad Javeriana Cali, physician with a master's degree in Epidemiology and a PhD in Health, for her valuable collaboration in the construction of the search equation and in the verification of relevant topics. Her contribution has been essential for the inclusion of key terms in this study, ensuring the quality and rigor of the work presented.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contribution statement

All the authors declare that the final version of this paper was read and approved. Authors and CRediT Roles: J.P.Y.:

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

Conceptualization, Investigation, Methodology, Data Curation, Formal Analysis, Software, Writing - Original Draft, Writing - Review & Editing, Visualization. A. F. P. M.: Investigation, Software, Writing - Original Draft, Visualization. D. F. M.: Supervision, Project administration. J. R. T. C: Supervision, Project administration. Writing – review & editing.

The total contribution percentage for this paper was as follows: J. P. Y.: 40 %, A. F. P. M.: 30 %, D. F. M.: 15 % and J. R. T. C.: 15 %.

Data availability statement

The data supporting the results of this study are only available for viewing GitHub through the link github.com/Felipep17/LDA-Dengue

Preprint

A preprint version of this article was deposited at: https://zenodo.org/records/13349632

References

Arenas Silva, Y. K. (2022). *Modelamiento de tópicos aplicado al análisis de contenido de los tweets sobre el dengue en Colombia* [Trabajo de grado] Universidad Industrial de Santander. Repositorio Institucional Noesis. https://noesis.uis.edu.co/server/api/core/bitstreams/5ea6ffd8-5471-4ceb-beef-4abbbe2244e9/content

Arteaga, D., & Mendoza, G. (2023). *Estudio sobre tendencias en líneas de investigación en los trabajos de grado del Programa de Estadística de la Universidad del Valle* [Trabajo de pregrado] Universidad del Valle. https://bibliotecadigital.univalle.edu.co/server/api/core/bitstreams/1d895306-087d-4b7d-aeac-89da3109a40b/content

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations. In *Advances in knowledge discovery and data mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010. Proceedings. Part I* (Vol. 14, pp. 391-402). Springer. https://doi.org/10.1007/978-3-642-13657-3_43

Asmussen, C. B., & Møller, C. (2019). Smart literature review: A practical topic modelling approach to exploratory literature review. *Journal of Big Data, 6*(93), 1-18. https://doi.org/10.1186/s40537-019-0255-7

Baum, D. (2012). Recognising speakers from the topics they talk about. *Speech Communication, 54*(10), 1132-1142. https://doi.org/10.1016/j.specom.2012.06.003

Benoit, K., Muhr, D., & Watanabe, K. (2021). *Stopwords: Multilingual stopword lists* (Version 2.3.0) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=stopwords

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(1), 993-1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*(1-9), 1775-1781. https://doi.org/10.1016/j.neucom.2008.06.011

Chatel-Chaix, L., Fischl, W., Scaturro, P., Cortese, M., Kallis, S., Bartenschlager, M., Fischer, B., & Bartenschlager, R. (2015). A combined genetic-proteomic approach identifies residues within Dengue virus NS4B critical for interaction with NS3 and viral replication. *Journal of Virology, 89*(14), 7170-7186. https://doi.org/10.1128/JVI.00867-15

Chiang, C., Beljanski, V., Yin, K., Olagnier, D., Ben Yebdri, F., Steel, C., Goulet, M. L., DeFilippis, V. R., Streblow, D. N., Haddad, E. K., et al. (2015). Sequence-specific modifications enhance the broad-spectrum antiviral response activated by RIG-I agonists. *Journal of Virology, 89*(15), 8011-8025. https://doi.org/10.1128/JVI.00845-15

da Silva Ferreira, E. R., de Oliveira Gonçalves, A. C., Tobal Verro, A., Undurraga, E. A., Lacerda Nogueira, M., Estofolete, C. F., & Santos da Silva, N. (2020). Evaluating the validity of dengue clinical-epidemiological criteria for diagnosis

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

in patients residing in a Brazilian endemic area. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 114*(8), 603-611. https://doi.org/10.1093/trstmh/traa031

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*(1), 61-84. https://doi.org/10.3166/dn.17.1.61-84

Díaz Rubiano, M. A. (2022). *Análisis de temas utilizando Twitter: Una aplicación del modelo LDA al caso colombiano* [Trabajo de grado] Universidad Santo Tomás. http://hdl.handle.net/11634/43303

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding. *Poetics, 41*(6), 570-606. https://doi.org/10.1016/j.poetic.2013.08.004

Elgesem, D., Feinerer, I., & Steskal, L. (2016). Bloggers' responses to the Snowden affair: Combining automated and manual methods in the analysis of news blogging. *Computer Supported Cooperative Work (CSCW), 25*(2-3), 167-191. https://doi.org/10.1007/s10606-016-9251-z

Elgesem, D., Steskal, L., & Diakopoulos, N. (2019). Structure and content of the discourse on climate change in the blogosphere: The big picture. En *Climate change communication and the internet* (pp. 21-40). Routledge. https://doi.org/10.1080/17524032.2014.983536

Evans, M. S. (2014). A computational approach to qualitative analysis in large textual datasets. *PloS One, 9*(2), e87908. https://doi.org/10.1371/journal.pone.0087908

Ghosh, D., & Guha, R. (2013). What are we 'tweeting' about obesity? Mapping tweets with topic modeling and Geographic Information System. *Cartography and Geographic Information Science, 40*(2), 90-102. https://doi.org/10.1080/15230406.2013.776210

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(S1), 5228-5235. https://doi.org/10.1073/pnas.0307752101

Gulo, C. A., & Rúbio, T. R. (2015). *Text mining scientific articles using R* [Ponencia]. Proceedings of the Doctoral Symposium in Informatics Engineering, Porto, Portugal. https://paginas.fe.up.pt/~prodei/dsie15/web/papers/dsie15_submission_10.pdf

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modeling. *Journalism & Mass Communication Quarterly, 93*(2), 332-359. https://doi.org/10.1177/1077699016639231

Guzman, M., Halstead, S., Art sob, H. *et al.* (2010). Dengue: A continuing global threat. *Nature Reviews Microbiology, 8*(12), S7-S16. https://doi.org/10.1038/nrmicro2460

Guzman-Ponce, A., Fernandez-Beltran, R., Valdovinos-Rosas, R. M., Romero-Huertas, M., & Marcial-Romero, J. R. (2023). Identification of latent topics in patients surviving COVID-19 in Mexico. *IEEE Latin America Transactions, 21*(3), 328-334. https://latamt.ieeer9.org/index.php/transactions/article/view/6995

Herbinger, K. H., Siess, C., Nothdurft, H., Von Sonnenburg, F., & Löscher, T. (2011). Skin disorders among travellers returning from tropical and non-tropical countries consulting a travel medicine clinic. *Tropical Medicine & International Health, 16*(11), 1457-1464. https://doi.org/10.1111/j.1365-3156.2011.02840.x

Iqtadar, S., Akbar, N., Huma, N., & Randhawa, F. A. (2017). Profile of hepatic involvement in dengue infections in adult Pakistani population. *Pakistan Journal of Medical Sciences, 33*(4), 963. https://doi.org/10.12669/pjms.334.13026

Jacobi, C., Van Atteveldt, W., & Welbers, K. (2018). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Rethinking research methods in an age of digital journalism. 4*(1) (pp. 89-106). Routledge. https://doi.org/10.1080/21670811.2015.1093271

Jiang, L. (2023). *Modelado de temas en documentos de texto: Análisis comparativo de LSA, PLSA y LDA* [Trabajo de fin de máster] Universitat Politècnica de València. Riunet. https://riunet.upv.es/handle/10251/197043

Jockers, M. L., & Mimno, D. (2013). Significant themes in 19th-century literature. *Poetics, 41*(6), 750-769. https://doi.org/10.1016/j.poetic.2013.08.005

Jones, C. H., Benítez-Valladares, D., Guillermo-May, G., Dzul-Manzanilla, F., Che-Mendoza, A., Barrera-Pérez, M., Selem-Salas, C., Chablé-Santos, J., Sommerfeld, J., Kroeger,

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470

A., O'Dempsey, T., Medina-Barreiro, A., & Manrique-Saide, P. (2014). Use and acceptance of long lasting insecticidal net screens for dengue prevention in Acapulco, Guerrero, Mexico. *BMC Public Health, 14*(1), 1-10. https://doi.org/10.1186/1471-2458-14-846

Kalra, V., Ahmad, S., Shrivastava, V., & Mittal, G. (2016). Quantitative and volume, conductivity and scatter changes in leucocytes of patients with acute undifferentiated febrile illness: A pilot study. *Transactions of The Royal Society of Tropical Medicine and Hygiene, 110*(5), 281-285. https://doi.org/10.1093/trstmh/trw028

Khetpal, A., Godil, A., Alam, M. T., Makhdoom, I. U. H. M., Adam, A. M., Mallick, A., Abbas, M. A., Abbas, A. H., Hasan, S. S., Shaikh, A., et al. (2021). Role of C-reactive proteins and liver function tests in assessing the severity of dengue fever. *JPMA. The Journal of the Pakistan Medical Association, 71*(3), 810-815. https://doi.org/10.47391/JPMA.170

Khurram, M., Qayyum, W., Umar, M., Jawad, M., Mumtaz, S., & Khaar, H. B. (2016). Ultrasonographic pattern of plasma leak in dengue haemorrhagic fever. *J Pak Med Assoc, 66*(2), 260-264.

Koltsova, O., & Koltcov, S. (2013). Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet, 5*(2), 207-227. https://doi.org/10.1002/1944-2866.POI331

Laoprasopwattana, K., Limpitikul, W., & Geater, A. (2020). Using clinical profiles and complete blood counts to differentiate causes of acute febrile illness during the 2009-11 outbreak of typhoid and chikungunya in a dengue endemic area. *Journal of Tropical Pediatrics, 66*(5), 504-510. https://doi.org/10.1093/tropej/fmaa006

Luquea, C., Rubriche, J., Galvis, J., & Sosa, J. (2021). Modelamiento de tópicos para identificar patrones en la investigación científica del Covid-19. *Comunicaciones en Estadística, 14*(1), 48-66. https://doi.org/10.15332/23393076.7705

Marks, M., Armstrong, M., Whitty, C. J., & Doherty, J. F. (2016). Geographical and temporal trends in imported infections from the tropics requiring inpatient care at the Hospital for Tropical Diseases, London-a 15-year study. *Transactions of the Royal Society of Tropical Medicine and Hygiene, 110*(8), 456-463. https://doi.org/10.1093/trstmh/trw053

Ministerio de Salud y Protección Social - Federación Médica Colombiana. (2013). *DENGUE- MEMORIAS, 2013. 2012-2013*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/TH/Memorias_dengue.pdf

Morizono, K., & Chen, I. S. (2014). Role of phosphatidylserine receptors in enveloped virus infection. *Journal of Virology, 88*(8), 4275-4290. https://doi.org/10.1128/JVI.03287-13

Nazareth, T., Teodósio, R., Porto, G., Gonçalves, L., Seixas, G., Silva, A. C., & Sousa, C. A. (2014). Strengthening the perception-assessment tools for dengue prevention: A cross-sectional survey in a temperate region (Madeira, Portugal). *BMC Public Health, 14*(39), 1-10. https://doi.org/10.1186/1471-2458-14-39

Nikita. (2020). *ldatuning: Tuning of the Latent Dirichlet Allocation models parameters* (R package version 1.0.2) [R package]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/package=ldatuning

Nivedita. (2016). Knowledge, attitude, behaviour and practices (KABP) of the community and resultant IEC leading to behaviour change about dengue in Jodhpur City, Rajasthan. *Journal of Vector Borne Diseases, 53*(4), 279-282.

Parkash, O., Almas, A., Jafri, S. W., Hamid, S., Akhtar, J., & Alishah, H. (2010). Severity of acute hepatitis and its outcome in patients with dengue fever in a tertiary care hospital Karachi, Pakistan (South Asia). *BMC Gastroenterology, 10*(1), 1-8.

Parra, D., Trattner, C., Gómez, D., Hurtado, M., Wen, X., & Lin, Y. R. (2016). Twitter in academic events: A study of temporal usage, communication, sentimental and topical patterns in 16 computer science conferences. *Computer Communications, 73*(1), 301-314. https://doi.org/10.1016/j.comcom.2015.07.001

Pattabhi, S., Wilkins, C. R., Dong, R., Knoll, M. L., Posakony, J., Kaiser, S., Mire, C. E., Wang, M. L., Ireton, R. C., Geisbert, T. W., et al. (2016). Targeting innate immunity for antiviral therapy through small molecule agonists of the RLR pathway. *Journal of Virology, 90*(5), 2372-2387. https://doi.org/10.1128/JVI.02202-15

Pilacuan-Bonete, L., Galindo-Villardón, P., & Delgado-Álvarez, F. (2022). HJ-Biplot as a tool to give an extra analytical boost for the Latent Dirichlet Allocation (LDA) model: With an application to digital news analysis about COVID-19. *Mathematics, 10*(14), 2529. https://doi.org/10.3390/math10142529

DOI: https://dx.doi.org/10.15359/ru.39-1.19
E-ISSN: 2215-3470
CC: BY-NC-ND

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science, 54*(1), 209-228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Rojas, R. M. R. (2022). Modelamiento de tópicos utilizando mensajes de Twitter relacionados al cáncer cervical. *Interfases, 16*(16), 41-52. https://doi.org/10.26439/interfases2022.n016.5887

Sahria, Y., & Fudholi, D. H. (2020). Analysis of health research topics in Indonesia using the LDA (latent Dirichlet allocation) topic modeling method. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi, 4*(2), 336-344. https://doi.org/10.29207/resti.v4i2.1821

Sanchez, L., Perez, D., Perez, T., Sosa, T., Cruz, G., Kouri, G., Boelaert, M., & Van Der Stuyft, P. (2005). Intersectoral coordination in Aedes aegypti control. A pilot project in Havana City, Cuba. *Tropical Medicine & International Health, 10*(1), 82-91. https://doi.org/10.1111/j.1365-3156.2004.01347.x

Schultes, O. L., Morais, M. H. F., Cunha, M. d. C. M., Sobral, A., & Caiaffa, W. T. (2021). Spatial analysis of dengue incidence and Aedes aegypti ovitrap surveillance in Belo Horizonte, Brazil. *Tropical Medicine & International Health, 26*(2), 237-255. https://doi.org/10.1111/tmi.13521

Sievert, C., & Shirley, K. (2015). *LDAvis: Interactive visualization of topic models* (Version 0.3.2) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=LDAvis

Silvestre Gómez, M. (2018). *Implementación de asignación jerárquica latente de Dirichlet para modelado de temas* [Trabajo de fin de máster, Universidad de Sevilla]. Archivo digital.

Thomas, L., Brouste, Y., Najioullah, F., Hochedez, P., Hatchuel, Y., Moravie, V., Kaidomar, S., Besnier, F., Abel, S., Rosine, J., Quenel, P., Césaire, R., & Cabié, A. (2010). Predictors of severe manifestations in a cohort of adult dengue patients. *Journal of Clinical Virology, 48*(2), 96-99. https://doi.org/10.1016/j.jcv.2010.03.008

Trueba-Gómez, R., & Estrada-Lorenzo, J. M. (2010). La base de datos PubMed y la búsqueda de información científica. *Seminarios de la Fundación Española de Reumatología, 11*(1), 49-63. https://doi.org/10.1016/j.semreu.2010.02.005

Van Atteveldt, W., Welbers, K., Jacobi, C., & Vliegenthart, R. (2014). LDA models topics... But what are "topics". In *Big Data in the Social Sciences Workshop*.

World Health Organization. (2023, December 21). *Disease Outbreak News; Dengue – Global situation*. https://www.who.int/emergencies/disease-outbreak-news/item/2023-DON498

Xie, X., Zou, J., Puttikhunt, C., Yuan, Z., & Shi, P. Y. (2015). Two distinct sets of NS2A molecules are responsible for dengue virus RNA synthesis and virion assembly. *Journal of Virology, 89*(2), 1298-1313. https://doi.org/10.1128/JVI.02882-14