Leveraging Topic Modelling to Analyze Biomedical Research Trends from the PubMed Database Using LDA Method

Yuri Pamungkas
Department of Medical Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
yuri@its.ac.id

***Abstract*—Biomedical research has become an essential entity in human life. However, finding trends related to research topics in the health sector contained in the repository is a challenging matter. In this study, we implemented topic modelling to analyze biomedical research trends using the LDA method. Topic modelling was carried out using data from 7000 articles from PubMed, which were processed with text processing such as lowercase, punctuation removal, tokenization, stop-word removal, and lemmatization. For topic modelling, the LDA with corpus conditions varied to 75% and 100% for validation. Alpha and beta parameters are also set with variations between 0.01, 0.31, 0.61, 0.91, symmetry, and asymmetry when the number of the corpus is changed. When the number of the corpus is 75%, the optimal number of topics is 7, with a coherence value of 0.52. Whereas when the number of the corpus is 100%, the optimal number of topics is 10 with a coherence value of 0.51. In addition, based on the results of article topic modelling, several topics are trending, including disease diagnosis, patient care, and genetic or cell research. Based on the classification of biomedical topics into seven categories, the optimal accuracy, precision, and recall values using the Random Forest algorithm were obtained, namely 85.57%, 87.36%, and 87.58%. The results of this study suggest that topic modelling using the LDA can be used to identify trends in biomedical research with high accuracy. This information can help stakeholders make informed decisions about the direction of future research.**

Keywords—Biomedical Research, Topic Modelling, LDA Method, PubMed, Text Processing

I. INTRODUCTION

Research in the health sector has become an essential entity in the academic world. The progress of this research can be seen in terms of the number of studies and the direction of the topics discussed [1]. The number of publications related to health research will continuously develop or increase in the journal repository [2]. Everyone (academicians, practitioners, and the general public) can easily search for or find scientific publications related to health research in the Open Journal System repository at PubMed. However, finding trends related to research topics in the health sector contained in the repository is a challenging matter [3]. Someone needs to read carefully one by one the various scientific publications that already exist. It will require a lot of effort and a lot of time. Therefore we need a method that aims to automate the reading of trends in research topics in the health sector in the PubMed repository. Automation in finding research topics can be done using a topic-modelling approach [4].

Topic modelling is an approach to analysing a collection of articles and grouping them into several clusters based on their level of similarity [5]. The purpose of topic modelling is to determine the topic automatically from a set of articles [6]. The researched articles have a hidden structure in the form of topics, distribution of topics per article, and determination of topics per word in each article [7]. Topic modelling uses a collection of these articles to infer the structure of hidden topics. In addition, the number of topics to be generated must also be determined before the topic modelling process is carried out [8]. The most commonly used topic modeling methods are Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis (PLSA) [9], which assume that each document consists of a mixture of hidden topics. Research in topic modeling includes developing more sophisticated models, such as models that take into account the hierarchical structure of issues, models that take into account time, or models that can handle text in different languages. This research has broad applications in various fields, including information management, sentiment analysis, recommendation systems, and social data analysis. With the continued development of technology and increasingly large amounts of data, research on topic modeling becomes increasingly important in understanding and exploring text information.

Generally, an article consists of many topics, and each topic consists of a distribution of words [10]. To group these topics, we need a method such as scoring the distribution of keywords (as identifiers for each topic) from related articles [11]. Therefore, LDA emerged as a solution to overcome this issue. Latent Dirichlet Allocation (LDA) is a generative model to find latent semantic topics in a text data set [12]. In addition, LDA is also a three-level hierarchical Bayesian model, where each item from a group of words is used to model a topic [13]. So, each topic is represented as a combination of the distribution of words that underlies the topic probability [14]. In the context of topic modelling, topic probabilities explicitly represent an article [15]. Latent Dirichlet Allocation also can overcome overfitting

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

problems that might occur in other types of methods [16]. Overfitting is a condition where the model is only suitable for specific samples. But in different samples, the match rate is low. Thus, the LDA method is considered ideal for exploring various topics in the article. Using Bayesian methods, LDA tries to find the most likely distribution of topics in a corpus based on the words observed in it. This process involves estimating the posterior distribution of topics within a text corpus based on a given initial distribution and information obtained from the words in each document. When the inference process is complete, LDA produces a representation of each document as a probability distribution over its possible topics. In addition, LDA also produces a representation of each topic as a probability distribution of the words that may appear in it. In this way, LDA allows us to understand the topics discussed in a text corpus and the words most likely associated with each of them.

Based on these, we propose the LDA topic modelling method to be implemented in the trend analysis of biomedical research topics. Topic modelling was carried out using data from 7000 articles. Scientific article data is taken from the PubMed database in modelling this research topic. After the article data is obtained, the next process is data processing using the Latent Dirichlet Allocation method to get topics often discussed in biomedical research. Knowledge regarding the trend of biomedical research topics is expected to be used as a reference in planning future national health research. In particular, to realize an efficient, quality, fair and sustainable national health system through an appropriate research ecosystem.

II. RELATED WORKS

Topic modelling is a development of text analysis that is useful in modelling textual data to find hidden topics. One model often used in topic modelling is Latent Dirichlet Allocation (LDA). The Latent Dirichlet Allocation (LDA) model is a probability model of textual data which can explain the correlation between words and semantic themes hidden in the document [20]. The parameter estimation used in the model is the Bayesian method. The Bayesian method is a method that provides estimation values through the posterior distribution. For this model, the calculation of the estimation of the posterior distribution is very complex, so Gibbs sampling estimation is used. Research related to topic modelling using the LDA method has also been carried out by many researchers. In their study, Gupta et al. [17] tried to predict research trends using LDA topic modelling. A total of 3269 research articles (with a span of 30 years) from the journal Applied Intelligence have been analyzed empirically. The method used to predict research topic trends is LDA with BoW (Bag of Words) and TF-IDF (Term Frequency-Inverse Document Frequency). The BoW calculates the frequency of word occurrences in all processed articles. Meanwhile, TF-IDF determines the relevance between keywords in several articles.

Based on the results of the predictions made in this study, the research topics developing the most rapidly from year to year are research related to AI-based algorithms, system optimization, and data processing. In their research, Rani et al. [18] tried to apply topic modelling to engineering and materials science articles. These fields were chosen because of their diverse and significant applications in various multidisciplinary disciplines, such as chemical engineering, industrial engineering, materials science, bioinformatics, and many more. By using the topic modelling method, grouping of keywords related to materials and material synthesis steps (such as centrifugation, milling, heating, and dissolving) can be carried out.

In addition, Choi et al. [19] attempted to examine research trends related to personal information using topic modelling. A total of 2356 research articles consisting of journals, conferences and book chapters with publication years ranging from 1972 to 2015 were further analyzed. The Latent Dirichlet Allocation method is used to identify topics from the abstracts of the articles. Based on research results for all types of articles, discussions on technology and privacy in social media (such as Facebook) have become prominent topics in the last few decades. However, this is different from the type of journal article which only focuses on health issues and digital business. It indicates that there are still considerable opportunities to explore research on personal information using topic modelling. Gupta et al. [21], in their research, predicted research/publication topics related to Covid-19 using Natural Language Processing (NLP) and Latent Dirichlet Allocation (LDA). In this study, Gupta et al. tried to cluster 25 topics from various publications in the LitCovid database (online literature focusing on health research). Based on the research that has been done, information is obtained that research topics related to Mental Health and Socio-Economic Impacts have increased. In contrast, research topics related to genome sequence have decreased. In addition, research topics related to Epidemiology tend to be constant. However, studies on clinical applications in health (such as personal protective equipment) and population-based epidemiology still need to be completed.

In their research, Kim et al. [22] combined a classification system based on ElasticSearch and topic modelling based on Latent Dirichlet Allocation to simplify searching for academic research results in online literature databases. As a result, several research topics, such as predictive analysis, learning processes, and data classification, are trending topics that often appear in current academic research. In addition, combining these methods (ElasticSearch and LDA) allows the estimation of keyword search results and paper research topics to shorten the processing time. This is because the proposed model can adequately analyze the correlation between words, the distribution of weights between words and their semantic similarities. In addition, Lee et al. [23] tried to analyze trends in sports research topics during Covid-19 using the LDA method. A total of 1604 sports-related research articles were collected and pre-processed to obtain abstract information and keywords. Then 17,018 keywords were extracted and weighted using the Term Frequency-Inverse Document Frequency method. This method is used to find out how often a word appears in the article. Extracted data were analyzed by topic modelling (LDA) and social network analysis. As a result, topics related to sports activities (with keywords such as games & Olympics), sports performance (with keywords such as adults & adolescents), and sports participation (with keywords such as metaverse & virtual) became quite trending research topics during Covid-19 takes place.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

III. METHODOLOGY

The following steps are carried out to implement LDA-based topic modeling in a collection of articles related to biomedical research.

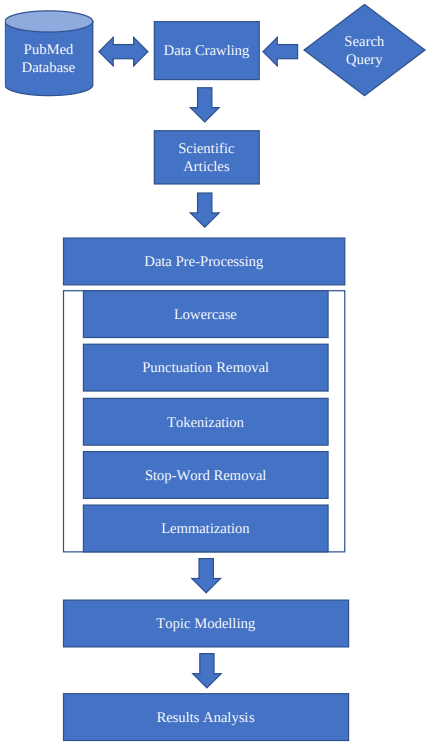

Fig. 1. Methodology

A. Data Collection

The articles used in this study were taken from the PubMed Database. PubMed is a literature search database under the auspices of the NCBI. In addition, PubMed is part of the United States National Library of Medicine (NLM) and a branch of the National Institutes of Health. PubMed comprises more than 30 million citations to biomedical literature (from MEDLINE, science journals, and online books). Such citations may include links to full-text content from PubMed Central and the publisher's website [24]. As many as 7,000 articles related to biomedical research (such as therapy, diagnostic procedures, preventive measures, diseases, medical records, biochemistry, human genome, and genetics) are stored in a CSV file. Each scientific article/paper contains a title, abstract, document text, and publication date.

B. Data Pre-Processing

The pre-processing stage is crucial in processing text data. Pre-processing aims to select and clean text data to make it even more structured. This stage consists of several processes: lowercasing, punctuation removal, tokenizing, stop-word removal, and lemmatization [25].

1. Lowercasing

In the initial preprocessing stage, the text will be changed from uppercase to lowercase so that the characters that will be accepted are only from "a" to "z." In this case, to avoid the existence of two identical words but considered different from the program because of differences in upper and lower case.

2. Punctuation Removal

Remove punctuation is a step taken to remove unused characters in the form of punctuation marks, numbers, markup/html/tags, and special characters ($, %, -, Etc). These characters may interfere with further processing and analysis of the text.

3. Tokenizing

Tokenizing is separating or cutting text according to its constituent words. This word separation aims to facilitate the following process: word counting, word weighting, and converting words into high-dimensional vectors.

4. Stop-Word Removal

Stop-words are words in documents with no meaning so that they can be omitted in text processing data analysis. Meanwhile, stop-word removal removes words with no meaning using the stop-list algorithm (removing less important words) or wordlist (saving essential words).

5. Lemmatization

Lemmatization is a process of finding the primary form of a word. This process aims to normalize text/words based on the primary form of the entry. Normalization can also be interpreted to identify and remove prefixes and suffixes from a word. Meanwhile, the lemma is the basic form of a word with a specific meaning based on the dictionary.

C. Topic Modelling

The topic is the distribution of some fixed vocabulary [26]. In simple terms, each document in the corpus contains its proportion of several topics discussed according to the words contained in the document [27]. Topic modelling consists of certain words that make up the topic, and a document has a certain probability of several topics being generated [28]. Alternatively, topic modelling is used to find the main topic hidden from a series of words in a large and unstructured document [29]. The topic modelling method analyzes data based on the original text regarding the relationship between topics with each other and the relationship between themes that can change at any time [30]. So, this method can be developed for searching or summarizing text contained in documents. In general, each document in the corpus will have a proportion of the topics covered. The results of the thematic modelling analysis will be used to summarize, visualize, explore and explain the information contained in the corpus [31]. Thus, topic modelling aims to find a collection of topics and words in various documents [32].

1. Latent Dirichlet Allocation

LDA (Latent Dirichlet Allocation) is the most popular topic modeling/analysis method for large unstructured documents. LDA can summarize, classify, relate, and process data because it generates several topics weighted in each document. LDA is an algorithm based on a probabilistic model with the assumption that each topic is a mixture of a collection of words and each document is a mixture of several topics [33]. The distribution

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

used to find the topic of a document is called the Dirichlet distribution. In addition, Dirichlet's results are used to allocate words to different topics. In LDA, documents are observed objects, while topics, topic distribution, and classification of each word on the topic are hidden structures, so this method is called Latent Dirichlet Allocation (LDA) [13].

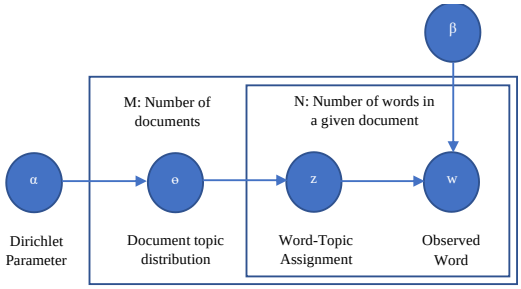

Fig. 2. LDA Architecture

The following is the formulation of the Latent Dirichlet Allocation (LDA) method [11].

$$P = \prod_{i=1}^{M} P(\theta_j; \alpha) \prod_{i=1}^{K} P(\emptyset; \beta) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) P(w_{j,t}|\emptyset z_{j,t}) \quad (1)$$

The α parameter is a Dirichlet distribution parameter that is useful for controlling the distribution of topics in each document, where the higher the α value indicates that the document contains mixed topics or several topics. Meanwhile, the lower value of α indicates that there are no topics mixed in the document. Therefore, a good distribution of topics can be seen from the lower α value. The β parameter is a Dirichlet distribution used to control the distribution of words on each topic, where the higher the β value indicates that the topic contains words found on other topics. Meanwhile, a low β value indicates that a topic consists of more specific words. The variable θ is a multinomial distribution or topic distribution for a particular document, which expresses the probability of a particular document. The higher the value of θ, the more topics are contained in the document. Meanwhile, the lower the value of θ, the more specific the number of topics in the document. The variable Z represents the topic for a particular word of a document. The variable w represents a specific word related to a certain topic in a document.

2. *Topic Coherence*

Topic coherence is a value used to evaluate topic modeling. With topic coherence, the semantic similarity between words in a document can be measured. These measurements assist in the process of distinguishing topics semantically from topics that are artifacts of statistical inference. In addition, topic coherence is considered capable of providing a better interpretation of the topic modeling results compared to Perplexity [12]. The division of the number of topics in the LDA process is essential in the topic modelling evaluation process. The results of determining the number of topics can be used to show the optimal model so that humans can understand it. One of the evaluation techniques for topic modelling is the Coherence score. The coherence score is a reference in determining the distribution of the optimal number of topics. The higher the coherence score on a certain number of k-topics, the more words can represent a particular topic.

$$C(w_i, w_j) = log \frac{P(w_i, w_j) + 1}{P(w_i).P(w_j)} \quad (2)$$

Where $C(w_i, w_j)$ is the coherence value, $P(w)$is the probability that the word (w) appears in the document, and $P(w_i, w_j)$ is the probability that the word (w_i, w_j) appears in the sliding window.

D. *Classification*

Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest are three popular classification algorithms in data analysis and machine learning. SVM is a classification algorithm that separates data classes by creating optimal decision lines or surfaces between these classes. The goal is to maximize the distance between the decision line and the closest data points from each class, called support vectors. SVM is suitable for complex or non-linear datasets because it can use kernel functions to transform data to higher dimensions to be separated linearly. KNN is a simple and intuitive classification algorithm. KNN uses the close distance between data points to determine their class. When given a new data point, the algorithm identifies the k nearest data points in the feature space (referred to as "neighbors") and takes the majority class of those neighbors as the predicted class of the new data point. Random Forest is an ensemble classification algorithm consisting of many decision trees. Each tree in a Random Forest is given a random sample from the training dataset and builds its model. When predicting the class for a new data point, each tree provides its class prediction, and the class most frequently chosen by those trees becomes the final prediction of the Random Forest.

These three algorithms have their respective strengths and weaknesses and are suitable for different datasets and classification problems. SVM is ideal for complex and well-structured datasets, and KNN is suitable for relatively small and not too complex datasets. At the same time, Random Forest is suitable for large datasets with many features and high complexity. In topic classification, all three can classify texts or documents into appropriate categories or topics.

IV. RESULTS AND DISCUSSIONS

In this study, we tried to carry out several variations in testing to get the best model for determining topics related to determining trends in biomedical research. So we varied the percentage of corpus numbers, alpha scores, and beta scores to get the best number of topics based on their coherence values [19]. The corpus is set to 75% and 100% for determining the number of topics [17]. In addition, alpha and beta parameters' values varied with values of 0.01, 0.31, 0.61, 0.91, symmetry and asymmetry.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

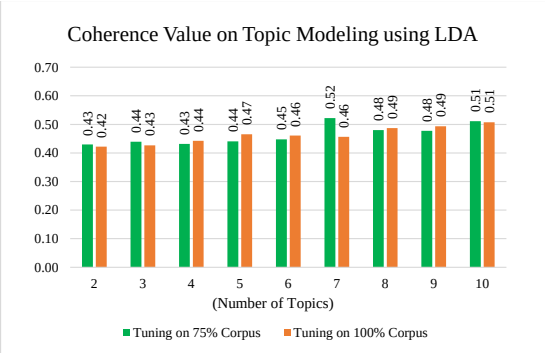

Fig. 3. Coherence Value on Topic Modeling using LDA

Based on model testing results using 75% corpus, the highest coherence value was obtained when the number of topics was 7 (coherence value 0.52). In testing the model using 100% corpus, the highest coherence value was obtained when the number of topics was 10 (coherence value 0.51).

TABLE I. TOPIC MODELING USING LDA (TUNING ON 75% CORPUS)

Validation Set	Topics	Alpha	Beta	Coherence
75% Corpus	7	0.01	0.01	0.398474
75% Corpus	7	0.01	0.31	0.384279
75% Corpus	7	0.01	0.61	0.395303
75% Corpus	7	0.01	0.91	0.451423
75% Corpus	7	0.01	symmetric	0.390551
75% Corpus	7	0.31	0.01	0.416892
75% Corpus	7	0.31	0.31	0.405067
75% Corpus	7	0.31	0.61	0.390503
75% Corpus	7	0.31	0.91	0.455646
75% Corpus	7	0.31	symmetric	0.417894
75% Corpus	7	0.61	0.01	0.407375
75% Corpus	7	0.61	0.31	0.429021
75% Corpus	7	0.61	0.61	0.381434
75% Corpus	7	0.61	0.91	0.44502
75% Corpus	7	0.61	symmetric	0.428097
75% Corpus	7	0.91	0.01	0.420782
75% Corpus	7	0.91	0.31	0.421603
75% Corpus	7	0.91	0.61	0.384744
75% Corpus	7	0.91	0.91	0.436384
75% Corpus	7	0.91	symmetric	0.402942
75% Corpus	7	symmetric	0.01	0.400696
75% Corpus	7	symmetric	0.31	0.393681
75% Corpus	7	symmetric	0.61	0.393842
75% Corpus	7	symmetric	0.91	0.454531
75% Corpus	7	symmetric	symmetric	0.414766
75% Corpus	7	asymmetric	0.01	0.404775
75% Corpus	7	asymmetric	0.31	0.379892
75% Corpus	7	asymmetric	0.61	0.40919
75% Corpus	7	asymmetric	0.91	0.522237
75% Corpus	7	asymmetric	symmetric	0.415347

In more detail, the coherence value of 0.52 at 75% of the corpus is obtained when the alpha value is "symmetry", and the beta value is 0.91. Furthermore, after obtaining the optimal number of topics (7 topics), keywords from each topic are extracted based on the weight of the words that appear in each document. Keywords such as "disease", "diagnosis", "treatment", "cell", and "gene" are some of the words that often appear in every document. It indicates that research topics such as disease diagnosis, patient care, and genetic/cell engineering are still the current research trends. Diagnosis of the disease is carried out to explain the clinical signs and symptoms experienced by a patient and distinguish it from other similar conditions [2]. While patient care or medical intervention is an attempt to cure, relieve, or control the disease, usually after the diagnosis of medical personnel. It can be surgical or non-surgical and generally involves some form of chemical or physical therapy [3]. In addition, research related to genes or cells can be used to predict health problems suffered and treatment of various hereditary diseases in humans [24].

TABLE II. THE DOMINANT KEYWORDS ON EACH TOPIC FROM A TOTAL OF 7 TOPICS

Topics	Keywords
Diagnosis and Treatment of Disease	case, disease, patient, treatment, clinical, diagnosis, tumor, report, present, use
Gene Study	cell, disease, protein, use, target, study, gene, expression, show, effect
Virus Infection and Mutation	infection, virus, cause, disease, isolate, variant, gene, strain, sample, mutation
Cause and Treatment of Disease	disease, treatment, induce, cell, case, liver, study, effect, increase, eye
Covid Study	study, disease, patient, health, use, covid, care, include, high, risk
High Risk Disease Study	patient, study, risk, disease, high, group, associate, use, year, analysis
Cell Study	cell, disease, level, study, expression, response, patient, associate, increase, induce

If the corpus value is changed to 100%, then the coherence value also tends to be constant (0.51) with an alpha "symmetry" value and a beta value of 0.91. Furthermore, after obtaining the optimal number of topics (10 topics), keywords from each topic are extracted based on the weight of the words that appear in each document.

TABLE III. TOPIC MODELING USING LDA (TUNING ON 100% CORPUS)

Validation Set	Topics	Alpha	Beta	Coherence
100% Corpus	10	0.01	0.01	0.407152
100% Corpus	10	0.01	0.31	0.393872

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

Validation Set	Topics	Alpha	Beta	Coherence
100% Corpus	10	0.01	0.61	0.453834
100% Corpus	10	0.01	0.91	0.465922
100% Corpus	10	0.01	symmetric	0.41205
100% Corpus	10	0.31	0.01	0.409806
100% Corpus	10	0.31	0.31	0.396801
100% Corpus	10	0.31	0.61	0.457528
100% Corpus	10	0.31	0.91	0.439195
100% Corpus	10	0.31	symmetric	0.410862
100% Corpus	10	0.61	0.01	0.418722
100% Corpus	10	0.61	0.31	0.396853
100% Corpus	10	0.61	0.61	0.443374
100% Corpus	10	0.61	0.91	0.433871
100% Corpus	10	0.61	symmetric	0.417585
100% Corpus	10	0.91	0.01	0.421449
100% Corpus	10	0.91	0.31	0.406688
100% Corpus	10	0.91	0.61	0.428761
100% Corpus	10	0.91	0.91	0.428833
100% Corpus	10	0.91	symmetric	0.416418
100% Corpus	10	symmetric	0.01	0.410043
100% Corpus	10	symmetric	0.31	0.387889
100% Corpus	10	symmetric	0.61	0.443027
100% Corpus	10	symmetric	0.91	0.455142
100% Corpus	10	symmetric	symmetric	0.403149
100% Corpus	10	asymmetric	0.01	0.428068
100% Corpus	10	asymmetric	0.31	0.419066
100% Corpus	10	asymmetric	0.61	0.444387
100% Corpus	10	asymmetric	0.91	0.507202
100% Corpus	10	asymmetric	symmetric	0.40161

Keywords such as "disease", "diagnosis", "treatment", "patient", and "genetic" are some of the words that often appear in every document. It indicates that even though the number of topic clusters has increased, research topics related to disease diagnosis, patient care, and genetics are still biomedical research trends based on our system's prediction results. In addition, the trend of keywords is almost the same when the number of topic clusters is 7 and 10, indicating that the proposed model tends to be stable for finding trends in biomedical research topics.

TABLE IV. THE DOMINANT KEYWORDS ON EACH TOPIC FROM A TOTAL OF 10 TOPICS

Topics	Keywords
Diagnosis and Treatment of Disease	case, disease, patient, treatment, clinical, diagnosis, report, present, lesion, surgery
Gene Study	disease, use, protein, base, target, method, drug, study, model, show
Virus Infection and Mutation	virus, infection, strain, viral, infect, isolate, test, sequence, positive, cause
Cause and Treatment of Disease	case, disease, cause, eye, induce, report, increase, day, effect, study
Covid Study	patient, disease, study, health, care, treatment, covid, include, clinical, use
High Risk Disease Study	patient, study, risk, group, high, disease, year, use, associate, analysis
Study of Infectious Diseases	covid, level, disease, infection, study, associate, patient, high, effect, risk
Clinical Study	model, use, disease, base, genetic, performance, clinical, study, brain, feature
Case Report	study, health, use, child, infection, prevalence, high, sample, report, risk
Cell Study	cell, disease, expression, gene, induce, study, protein, role, treatment, cancer

In classifying biomedical research topics into seven topics, this research uses several algorithms such as SVM, K-NN, NB (Naïve Bayes), and RF (Random Forest). Model evaluation is carried out by dividing training and testing data into several ratios of 60:40, 70:30, and 80:20. The model testing results are observed through the parameters of accuracy, precision, and recall. The following are the results of the model evaluation.

TABLE V. EVALUATION RESULTS

Algorithm	Evaluation Parameter	Results (%)		
		(60:40)	(70:30)	(80:20)
SVM	Accuracy	68.57	82.38	78.57
	Precision	71.08	82.30	79.10
	Recall	74.68	84.55	85.89
K-NN	Accuracy	70.71	83.81	85.00
	Precision	73.05	84.07	86.42
	Recall	76.73	85.59	87.50
NB	Accuracy	66.61	82.34	77.14
	Precision	69.97	82.96	77.97
	Recall	71.52	83.71	84.66
RF	Accuracy	82.86	84.29	85.57
	Precision	79.10	84.89	87.36
	Recall	92.72	85.65	87.58

V. CONCLUSSIONS

Topic modelling is a method to find the main topic hidden from a series of words in a large and unstructured document. In

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

this study, the documents analyzed were articles related to biomedical research from the PubMed Database. As many as 7,000 articles related to biomedical research were processed using text processing using the LDA method.

The stages in the initial processing of text data include Removing punctuation, Tokenizing, Stopwords, and Lemmatization. These stages are crucial because they aim to select and clean text data so that it becomes even more structured. Then, the LDA method was used for topic modelling with corpus conditions ranging from 75% to 100% for the validation process. Alpha and beta parameters are also set with variations between 0.01, 0.31, 0.61, 0.91, symmetry, and asymmetry when the number of the corpus is changed. When the number of the corpus is 75%, the optimal number of topics is 7 clusters with a coherence value of 0.52. Whereas when the number of the corpus is 100%, the optimal number of topics is 10 clusters with a coherence value of 0.51.

In addition, based on the results of the implementation of topic modelling in biomedical-related research articles, several topics are currently being researched, including disease diagnosis, patient care, and genetic or cell research. Based on the classification of biomedical topics into seven categories, the optimal accuracy, precision, and recall values using the Random Forest algorithm were obtained, namely 85.57%, 87.36%, and 87.58%. The results of this study suggest that topic modelling using the LDA can be used to identify trends in biomedical research with high accuracy. This information can help stakeholders make informed decisions about the direction of future research. For further research, we will increase the number of biomedical research articles used in topic modelling and pay more attention to the computational time of the LDA method in clustering document topics (because this is a limitation of this research). It aims to optimize the model further to make the results more accurate in modelling research trends.

REFERENCES

[1] Kesiku, C.Y., Chaves-Villota, A., & Garcia-Zapirain, B. (2022). Natural Language Processing Techniques for Text Classification of Biomedical Documents: A Systematic Review. Information. 13(10):49. https://doi.org/10.3390/info13100499

[2] Capuano, N., Foggia, P., Greco, L., & Ritrovato, P. (2022). A Linked Data Application for Harmonizing Heterogeneous Biomedical Information. Appl. Sci. 12(18):9317. https://doi.org/10.3390/app12189317

[3] Chaves, A., Kesiku, C., & Garcia-Zapirain, B. (2022). Automatic Text Summarization of Biomedical Text Data: A Systematic Review. Information. 13(8):393. https://doi.org/10.3390/info13080393

[4] Pedral Sampaio, R., Aguiar Costa, A., & Flores-Colen, I. (2022). A Systematic Review of Artificial Intelligence Applied to Facility Management in the Building Information Modeling Context and Future Research Directions. Buildings. 12(11):1939. https://doi.org/10.3390/buildings12111939

[5] Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R., & Sattar, A. (2023). Topic Classification of Online News Articles Using Optimized Machine Learning Models. Computers. 12(1):16. https://doi.org/10.3390/computers12010016

[6] Asmussen, C.B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data. 6 (93) . https://doi.org/10.1186/s40537-019-0255-7

[7] Albalawi, R., Yeap, T.H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. Front. Artif. Intell. 3:42. doi: 10.3389/frai.2020.00042

[8] Silva, C.C., Galster, M. & Gilson, F. (2021). Topic modeling in software engineering research. Empir Software Eng. 26:120. https://doi.org/10.1007/s10664-021-10026-0

[9] Owa, D. (2021) Identification of Topics from Scientific Papers through Topic Modeling. Open Journal of Applied Sciences, 11, 541-548. doi: 10.4236/ojapps.2021.104038.

[10] Kukushkin, K., Ryabov, Y., & Borovkov, A. (2022). Digital Twins: A Systematic Literature Review Based on Data Analysis and Topic Modeling. Data. 7(12):173. https://doi.org/10.3390/data7120173

[11] Chiu, C.-C., Wu, C.-M., Chien, T.-N., Kao, L.-J., & Qiu, J.T. (2022). Predicting the Mortality of ICU Patients by Topic Model with Machine-Learning Techniques. Healthcare. 10(6):1087. https://doi.org/10.3390/healthcare10061087

[12] Wu, F., Xu, W., Lin, C., & Zhang, Y. (2022). Knowledge Trajectories on Public Crisis Management Research from Massive Literature Text Using Topic-Clustered Evolution Extraction. Mathematics. 10(12):1966. https://doi.org/10.3390/math10121966

[13] Delcea, C., Cotfas, L.-A., Crăciun, L., & Molănescu, A.G. (2022). New Wave of COVID-19 Vaccine Opinions in the Month the 3rd Booster Dose Arrived. Vaccines. 10(6):881. https://doi.org/10.3390/vaccines10060881

[14] Ong, S.-Q., Pauzi, M.B.M., & Gan, K.H. (2022). Text Mining and Determinants of Sentiments towards the COVID-19 Vaccine Booster of Twitter Users in Malaysia. Healthcare. 10(6):994. https://doi.org/10.3390/healthcare10060994

[15] Feizollah, A., Anuar, N.B., Mehdi, R., Firdaus, A., & Sulaiman, A. (2022). Understanding COVID-19 Halal Vaccination Discourse on Facebook and Twitter Using Aspect-Based Sentiment Analysis and Text Emotion Analysis. Int. J. Environ. Res. Public Health. 19(10):6269. https://doi.org/10.3390/ijerph19106269

[16] Gourisaria, M.K., Chandra, S., Das, H., Patra, S.S., Sahni, M., Leon-Castro, E., Singh, V., & Kumar, S. (2022). Semantic Analysis and Topic Modelling of Web-Scrapped COVID-19 Tweet Corpora through Data Mining Methodologies. Healthcare. 10(5):881. https://doi.org/10.3390/healthcare10050881

[17] Gupta, R. K., Agarwalla, R., Naik, H. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of research trends using LDA based topic modeling. Global Transitions Proceedings. 3:1 (298-304). https://doi.org/10.1016/j.gltp.2022.03.015

[18] Rani, S., & Kumar, M. (2021). Topic modeling and its applications in materials science and engineering. Materialstoday:Proceedings. 45:6 (5591-5596). https://doi.org/10.1016/j.matpr.2021.02.313

[19] Choi, H. S., Lee, W. S., & Sohn, S. Y. (2017). Analyzing research trends in personal information privacy using topic modeling. Computers & Security. 67 (244-253). https://doi.org/10.1016/j.cose.2017.03.007

[20] Trivedi, S.K., Patra, P., Singh, A., Deka, P. and Srivastava, P.R. (2022), "Analyzing the research trends of COVID-19 using topic modeling approach", Journal of Modelling in Management, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JM2-02-2022-0045

[21] Gupta, A., Aeron, S., Agrawal, A., & Gupta, H. (2021). Trends in COVID-19 Publications: Streamlining Research Using NLP and LDA. Front. Digit. Health 3:686720. doi: 10.3389/fdgth.2021.686720

[22] Kim, M., & Kim, D. A. (2022). Suggestion on the LDA-Based Topic Modeling Technique Based on ElasticSearch for Indexing Academic Research Results. Appl. Sci. 12, 3118. https://doi.org/10.3390/app12063118

[23] Lee, J. W., Kim, Y. B., & Han, D. H. (2022). LDA-based topic modeling for COVID-19 related sports research trends. Front. Psychol. 13:1033872. doi: 10.3389/fpsyg.2022.1033872

[24] Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Smart Wearables for the Detection of Cardiovascular Diseases: A Systematic Literature Review. Sensors. 23(2):828. https://doi.org/10.3390/s23020828

[25] Davagdorj, K., Wang, L., Li, M., Pham, V.-H., Ryu, K.H., & Theera-Umpon, N. (2022). Discovering Thematically Coherent Biomedical Documents Using Contextualized Bidirectional Encoder Representations from Transformers-Based Clustering. Int. J. Environ. Res. Public Health. 19(10):5893. https://doi.org/10.3390/ijerph19105893

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024

[26] Diaf, S., & Fritsche, U. (2022). Topic Scaling: A Joint Document Scaling–Topic Model Approach to Learn Time-Specific Topics. Algorithms. 15(11):430. https://doi.org/10.3390/a15110430

[27] Duan, Z., Lu, L., Yang, W., Wang, J., & Wang, Y. (2022). An Abstract Summarization Method Combining Global Topics. Appl. Sci. 12(20):10378. https://doi.org/10.3390/app122010378

[28] Scarpino, I., Zucco, C., Vallelunga, R., Luzza, F., & Cannataro, M. (2022). Investigating Topic Modeling Techniques to Extract Meaningful Insights in Italian Long COVID Narration. BioTech. 11(3):41. https://doi.org/10.3390/biotech11030041

[29] Xu, H., Zhang, M., Zeng, J., Hao, H., Lin, H.-C.K., & Xiao, M. (2022). Use of Latent Dirichlet Allocation and Structural Equation Modeling in Determining the Factors for Continuance Intention of Knowledge Payment Platform. Sustainability. 14(15):8992. https://doi.org/10.3390/su14158992

[30] Hananto, V.R., Serdült, U., & Kryssanov, V. (2022). A Text Segmentation Approach for Automated Annotation of Online Customer Reviews, Based on Topic Modeling. Appl. Sci. 12(7):3412. https://doi.org/10.3390/app12073412

[31] Murakami, R., & Chakraborty, B. (2022). Investigating the Efficient Use of Word Embedding with Neural-Topic Models for Interpretable Topics from Short Texts. Sensors. 22(3):852. https://doi.org/10.3390/s22030852

[32] Sangaji, A. H., Pamungkas, Y., Nugroho , S. M. S. ., & Wibawa , A. D. (2022). Rule-based Disease Classification using Text Mining on Symptoms Extraction from Electronic Medical Records in Indonesian. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 7(1), 69-80. https://doi.org/10.22219/kinetik.v7i1.1377

[33] Quatrini, E., Colabianchi, S., Costantino, F., & Tronci, M. Clustering Application for Condition-Based Maintenance in Time-Varying Processes: A Review Using Latent Dirichlet Allocation. Appl. Sci. 12(2):814. https://doi.org/10.3390/app12020814

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i2.2117, Copyright ©2024
Submitted : March 7, 2024, Revised : April 3, 2024, Accepted : April 4, 2024, Published : June 15, 2024