Bulletin of Electrical Engineering and Informatics
Vol. 15, No. 6, December 2026, pp. 5556~5568
ISSN: 2302-9285, DOI: 10.11591/eei.v15i6.14194

Grid search-based LDA optimization for agricultural topic discovery and intertopic distance map visualization

Maukar[1], Diana Tri Susetianingtias[2], Eka Patriya[3], Rodiah[4]
[1]Graduate Program, Master of System Information Management, Gunadarma University, Depok, Indonesia
[2]Department of Computer System, Faculty of Informatic System and Information Technology, Gunadarma University, Depok, Indonesia
[3]Department of Management, Faculty of Economy, Gunadarma University, Depok, Indonesia
[4]Department of Informatics, Faculty of Industry Technology, Gunadarma University, Depok, Indonesia

Article Info

Article history:

Received Aug 8, 2026
Revised Sep 8, 2026
Accepted Sep 12, 2026

Keywords:

Agricultural research topics
Big scholarly data
Grid search
Intertopic distance map
Latent Dirichlet allocation

ABSTRACT

Agricultural research has generated a substantial volume of scientific publications, yet this corpus remains underutilized for systematically identifying major research topics and thematic structures. This study aims to identify dominant topics in agricultural publications from 2012 to 2023 using latent Dirichlet allocation (LDA) with systematic hyperparameter tuning and intertopic distance map visualization. An exploratory quantitative approach was applied to 75,031 publications retrieved from Scopus, ProQuest, and Web of Science (WoS). The analysis involved text preprocessing, phrase modeling, dictionary and corpus construction, and LDA topic modeling. An exhaustive grid search over a discretized parameter space identified a 17 topic model with alpha=0.1, beta=0.70, and a coherence score of 0.5075, followed by interactive visualization using pyLDAvis. The results revealed six major thematic groups: agricultural practices and food systems; data analysis applications; plant production and growth factors; environmental and land-use changes; effects and sample analysis; and plant genetics under environmental pressures. The findings demonstrate that LDA combined with systematic parameter tuning can reveal latent thematic structures and clarify relationships among agricultural research topics. The study provides a data-driven basis for exploratory mapping of research directions, while temporal trend evolution and direct policy implications require further longitudinal analysis.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Maukar
Graduate Program, Master of System Information Management, Gunadarma University
Margonda Raya 100, Pondok Cina, Depok, West Java, Indonesia
Email: maukar@staff.gunadarma.ac.id

1. INTRODUCTION

From an information-processing standpoint, the central challenge addressed in this study is how to transform a large, heterogeneous, multi-source corpus of unstructured scientific text into a scalable and navigable representation of latent themes, with agriculture serving as the application domain. Agriculture plays an important role in global food security [1], the provision of industrial raw materials, and employment. As food demand and environmental challenges increase, the sector is undergoing significant transformation through technology adoption [2], data-driven innovation, and precision farming practices [3]. This transformation is accompanied by a rapid growth of agricultural scientific publications covering diverse issues, including soil quality, crop resilience, sensor technology, artificial intelligence, and climate change [4]. The increasing volume and heterogeneity of these publications make it difficult to identify dominant research trends, research development patterns, and strategic directions [5].

To address these challenges, natural language processing (NLP) and topic modeling have become effective approaches for extracting thematic structures from unstructured text [6]. Latent Dirichlet allocation (LDA) [7] is one of the most widely used unsupervised probabilistic methods for uncovering latent themes in large document collections [8]. Previous studies have applied LDA to various domains. Nguyen *et al*. [9] examined LDA for large and unstructured text datasets and demonstrated its ability to identify hidden patterns and thematic structures, while also highlighting its sensitivity to the number of topics and parameter settings.

Kallens and Christiansen [10] applied LDA to 102 scientific publications on mathematical creative thinking (MCT) in Indonesia collected from the Garuda database for the period 2010–2022. The analysis identified 23 topics representing major aspects of MCT research, including learning, skills, assessment, motivation, and teaching models. Similarly, Mao *et al*. [11] analyzed 4,327 article abstracts from Scopus on organic agriculture and entrepreneurship for the period 1946–2023. The study identified five dominant topics related to soil, environment, farming business, organic production, and organic substances, with a coherence score of 0.419.

Ogunleye *et al*. [12] integrated LDA-based topic modeling with classification algorithms, including k-NN, support vector machine (SVM), Naive Bayes, and decision trees (DT), for digital content analysis. The combined approach improved classification performance, with the DT achieving an accuracy of 98.86%. Setiawan *et al*. [13] analyzed 16,039 Scopus abstracts published between 1977 and 2023 to examine the development of global bibliometric research. Using LDA together with word embedding and topic-number selection, the study identified four major topics and examined changes in the conceptual, intellectual, and social structures of bibliometric research.

Various studies have therefore demonstrated the effectiveness of LDA in identifying latent thematic structures across different research domains [10]–[12]. However, the existing studies generally rely on a single data source, use fixed or informally selected hyperparameters, or provide limited integration of cross-topic thematic visualization using intertopic distance maps. In addition, few studies systematically combine multi-source scholarly data at scale to investigate agricultural research themes over an extended period. Table 1 summarizes the characteristics, data sources, methodological approaches, and key differences between the related LDA-based studies and the present study.

Table 1. Comparison of related LDA-based topic-modeling studies

Study	Dataset/source	Preprocessing	Parameter selection	Evaluation metric	Visualization
[9]	Illustrative real-world text corpora (multi-domain).	Not detailed.	Not systematically searched; parameters not reported as optimized.	Qualitative interpretability.	Not reported.
[10]	102 articles, Garuda database, 2010-2022 (MCT and Indonesia).	Text cleaning.	Fixed configuration (VEM); no parameter search reported.	Qualitative topic labeling.	Not reported.
[11]	4,327 abstracts, Scopus, 1946-2023 (organic agriculture).	Not detailed.	Fixed configuration; no parameter search reported.	C_v coherence=0.419.	Not reported.
[12]	Large-scale text data (digital content reviews).	Not detailed.	Not the focus of the study (classifier tuning only).	Classification accuracy (e.g., DT=98.86%).	Not reported.
[13]	16,039 abstracts, Scopus, 1977-2023 (bibliometrics).	Not detailed.	Not systematically searched.	Qualitative+word embedding exploration.	Word cloud, dynamic exploration.
This study	75,031 abstracts, Scopus+ProQuest+Web of Science (WoS), 2012-2023 (agriculture).	Case folding, bigram/trigram phrase modeling, stopword removal, lemmatization.	Exhaustive grid search: 800 (topic number (k), alpha (α), and beta (β)) x2 corpus-subset configurations.	C_v coherence=0.5075 (final model).	Intertopic distance map (pyLDAvis).

As Table 1 shows, prior LDA-based studies [9]–[13] generally relied on a single data source and fixed or informally selected hyperparameters, with limited reporting of systematic parameter search, computational cost, and integrated intertopic distance map visualization. The present study addresses this gap by combining a multi-source corpus, exhaustive LDA hyperparameter search, and intertopic distance map visualization within a reproducible workflow rather than proposing a new topic-modeling algorithm. The problem is fundamentally an information-processing task: transforming a large, heterogeneous, unstructured scholarly corpus into a structured representation of latent themes. This informatics contribution lies in the reproducible computational workflow, combinatorial hyperparameter optimization, and interactive visualization for automated knowledge discovery.

Previous studies such as [10]–[13] commonly used fixed or informally selected LDA parameters without systematically examining the effects of k, α, β, preprocessing, and corpus scale. This study therefore

treats hyperparameter selection as a combinatorial search problem, jointly evaluating k, α, and β across 75% and 100% corpus sizes using the C_v coherence score. The corpus comprises 75,031 publications from Scopus, ProQuest, and WoS covering 2012–2023. LDA is used to identify latent topics, while intertopic distance map visualization is applied to examine their semantic relationships. Temporal evolution is not claimed as a finding because the analysis uses an aggregated corpus.

This study is guided by the following research questions:

RQ1. How do the k, α, and β jointly affect the C_v coherence of an LDA model trained on a large, multi-source agricultural corpus?

RQ2. Does an exhaustive grid search over these hyperparameters yield a more coherent topic solution than an untuned, default configuration?

RQ3. Does an intertopic distance map improve the interpretability of the resulting topic structure relative to ranked topic-word lists alone?

2. METHOD

From an information-system perspective, the research method consists of seven interacting stages; i) data acquisition from Scopus, ProQuest, and WoS, ii) data cleaning and attribute standardization, iii) NLP preprocessing, including case folding, tokenization, stopword removal, and lemmatization, iv) phrase-based feature construction using bigram and trigram modeling, v) numerical feature representation through dictionary and bag-of-words (BoW) corpus construction, vi) combinatorial hyperparameter optimization and LDA-based topic modeling, and vii) topic evaluation and intertopic distance map visualization. This workflow is designed as a reproducible information-processing procedure for large-scale scholarly text.

The study analyzes agricultural publications from 2012–2023 collected from Scopus, ProQuest, and WoS using the keyword "Agriculture" through database search and export procedures. During preprocessing, case folding and the removal of punctuation, special characters, and unnecessary whitespaces are performed, followed by tokenization, stopword removal, and lemmatization. Bigram and trigram models are then constructed to capture multi-word expressions (MWEs) before the text is converted into numerical representations for LDA modeling. The LDA modeling stage consists of two models. The base model is used to calculate topic coherence and keyword weights during the hyperparameter evaluation process. The final model is subsequently trained using the optimal parameters obtained from the combinatorial hyperparameter search and used for topic interpretation and intertopic distance map visualization. The complete research workflow is presented in Figure 1.

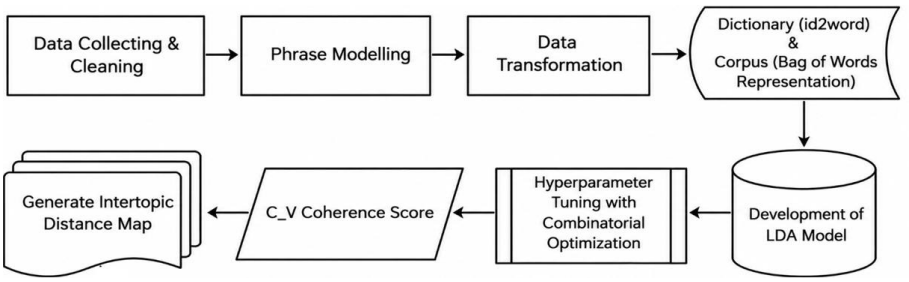


Figure 1. Stages of combinatorial modeling research using LDA

2.1. Data collecting and cleaning

Because the corpus was assembled from three databases with different export formats, field conventions, and metadata structures, data collection was treated as an interoperability and data-standardization process. Records from Scopus, ProQuest, and WoS were retrieved using the single keyword "*Agriculture*" through institutional access and exported in *.xlsx* format. Each dataset was converted to a common *.csv* schema containing four harmonized attributes, author, title, year, and abstract. The standardized datasets were then merged and cleaned, resulting in a total of 75,031 records.

2.2. Preprocessing

This study uses textual data from abstracts of scientific publications as input to the LDA model, therefore before the preprocessing stage, attributes in the dataset other than the '*Abstract*' attribute are removed so that preprocessing is focused on the words in the abstract of the scientific publication. The preprocessing stage includes several steps, namely:

2.2.1. Phrase modeling

Phrase modeling is used to detect and form multi-word expressions [10] or commonly occurring phrases in the text. This step is important in topic modeling because LDA treats words as individual units, although many meaningful concepts are expressed as phrases [11], [12].

a. Bigram and trigram detection, a bigram consists of two adjacent words that occur frequently enough to form a semantic unit [13], while a trigram consists of three consecutive words that frequently occur together [14], [15]. In this study, the Gensim Phrases class is used to detect bigrams and trigrams based on word co-occurrence statistics [16]. The association score between two words is calculated using (1):

$$Score(W_i, W_j) = \frac{(count(w_i,w_j) - \min_count)}{countw_i \times count(w_j)} \times N \quad (1)$$

The association score is determined from *count (wi, wj)*, representing the number of occurrences of words *wi* and *wj* in sequence, *count(wi)* and *count(wj)*, representing their individual frequencies, and *N*, representing the total number of words in the corpus. The *min_count* parameter defines the minimum frequency required for a word pair to be considered, while *threshold* defines the minimum association score for forming a phrase. If the association score exceeds the threshold, the words are merged using an underscore (_) as a single phrase unit. In this study, the Gensim *Phrases* class uses *min_count*=5 and *threshold*=10.0 for both bigram and trigram detection. No corpus-specific tuning of these parameters was performed.

b. Stopword removal, stopword removal is performed using the English stopword list provided by the NLTK library because the dataset consists of English-language scientific publications [17], [18]. Common words such as "*the*", "*is*", "*and*", and "*of*" are removed because they provide limited semantic information and may introduce noise. This step reduces data dimensionality and supports improved topic coherence in LDA [19], [20].

c. Perform lemmatization to change a word to its basic form (lemma) [16], for example, "*running*" becomes "*run*" and "*studies*" becomes "*study*". This study performs lemmatization using spaCy, with part-of-speech information used to retain nouns, adjectives, verbs, and adverbs. The resulting lemmas are used as the normalized lexical units for subsequent dictionary and corpus construction.

$$P(y_i|x) = \frac{1}{Z(x)} \exp\left(\sum_j \lambda_j f_j(y_{i-1}, y_i, x, i)\right) \quad (2)$$

In (2) is used to calculate the probability of the label (tag) sequence to-*i* that is y_i taking into account the sequence of inputs *x* (e.g., a word in a sentence), and the preceding label where this equation is a sequential model that considers the context between the labels. Morpho-lexical rules are linguistic rules that combine morphology, namely: word structure, such as suffixes (prefixes, suffixes), plural forms, verb conjugations, and lexicons, namely: dictionary of root words (lemma) and their derivative forms [21], [22]. For example, "*running*" has the lemma "*run*" by removing the suffix "*-ing*", while the irregular comparative form "*better*" has the lemma "*good*" according to a specific morpho-lexical rule.

2.2.2. Data transformation for latent Dirichlet allocation

At this stage, the text data that has gone through the process Lemmatization converted to numerical representations so that they can be processed by the LDA algorithm [16]. This process is carried out in two main steps:

a. Dictionary construction, this research uses the Gensim library Corpora. Dictionary, which is a structure that maps each unique word in a document to a numeric ID [16], [23]. The goal is to build a dictionary with the format Where is the word to that the i word has been dilemmatized and is the unique numerical ID for the word.

b. Corpus construction BoW representation), once the dictionary is created, each document in the corpus will be converted into a list of pairs Where denotes the ID of the *i* word in the dictionary, and frequency is the frequency with which the word appears in the document. Mathematically, the process of transforming this text data is:

– Mapping words to ID with is V vocabulary.
– The representation of the document where each document d is represented with where *i* is the ID of the *i*-word and is the frequency of the *i*-word in document *d*.
– A corpus is expressed as the sum of all representations of documents in pairs (*word_id*, frequency).

2.3. Development of latent Dirichlet allocation model

This stage aims to build the base model LDA, which is the initial model used to find the distribution of topics in a corpus of documents and to calculate topic coherence [24]–[26], a measure of how consistent or interpretable the resulting topics are. This model is built using *LdaModel()* from the *gensim.models.ldamodel* module and requires several common parameters, as shown in Table 2.

Table 2. Parameters of LDA model

Parameters	Function
corpus'	Stores the corpus to be used in the LDA model.
'id2word'	Stores the dictionary containing the word ID mapping, used so the LDA model can interpret word IDs in the corpus.
'num_topics'	Declares the number of topics that the LDA model will find.
'random_state'	Ensures the topics generated by the LDA model are consistent every time the model is run using the same corpus and dictionary.
'chunksize'	Manages the number of documents the model processes at one time.
'passes'	States the number of times the entire corpus passes through the model during training.
'alpha'	Hyperparameter affecting the distribution of topics in a document (Dirichlet distribution parameter for topic distribution).
'eta'	Hyperparameter affecting the distribution of words in a topic (Dirichlet distribution parameter for word distribution).
'iterations'	Specifies the number of iterations performed on each document to find its topic distribution.
'per_word_topics'	Allows the model to track which topics each word is most likely associated with.

Mathematically, LDA models the generative process of documents as follows [27]–[29]:

a. For each document $d \in D$ then take the topic distribution $\theta_d \sim Dir(\alpha)$.
b. For each w word in document d, then:

– Take topics $Z_{d,n} \sim Multinomial(\theta_d)$
– Take the word $w_{d,n} \sim Multinomial(\phi_{z_{d,n}})$, with $\phi_k \sim Dir(\beta)$

The combined probability equation of the LDA model uses in (3):

$$P(D, Z, \theta|\alpha, \beta) = \prod_{d=1}^{M} P(\theta_d|\alpha) \prod_{k=1}^{k} P(\phi_k|\beta) \prod_{n=1}^{N_d} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{z_{d,n}}) \quad (3)$$

The combined probability equation in the LDA model expresses the entire generative process of a text document based on the distribution of topics and words. The LDA model assumes that each document D it is formed through a probabilistic process based on two main parameters, i.e., α and β, which control the distribution of topics in documents and the distribution of words in topics [18], [30]. This combined equation calculates the shared probabilities of the entire corpus of documents D, topic-by-document distribution θ, word-by-topic distribution ϕ and latent topic variables z allocated to each word. The first component $\prod_{d=1}^{M} P(\theta_d|\alpha)$ describes the probability that the document d has a topic distribution θ_d derived from a Dirichlet distribution with α parameter. The second component $\prod_{k=1}^{k} P(\phi_d|\beta)$ shows that each k topic has a ϕ_k word distribution that also follows the Dirichlet distribution with the β parameter. Furthermore, for each word n in the d document, the component $(z_{d,n}|\theta_d)$ represents the probability that the word is allocated to a particular topic, whereas $P(w_{d,n}|\phi_{z_{d,n}})$ describe the probability of the word appearing based on the distribution of the word from the selected topic. This combined equation illustrates the overall probability of a set of documents resulting from topic allocation and word selection based on hidden distributions. The main purpose of inference in LDA is to find the best estimates of the two hidden distributions, θ, i.e., the distribution of topics in each document, and ϕ, i.e., the distribution of words in each topic, based on the observed documents D. This process enables the model to identify the latent (not directly visible) thematic structure of a set of text documents. For notational clarity, this study uses β and eta (η) interchangeably to denote the Dirichlet prior over the topic-word distribution φ: eta is the parameter name used in the Gensim implementation (Table 1), while β is the conventional statistical notation used in the LDA generative model of in (3).

The BoW corpus and dictionary are used to train the LDA base model with the specified k, α, β, chunk size, number of passes, iterations, and random state. The model estimates topic distributions for documents and word distributions for topics using Dirichlet priors. After training, preliminary trials were conducted to determine suitable values of kk and chunk size for subsequent hyperparameter tuning. The trials used kk values of 5, 10, 15, and 20 and chunk sizes of 10, 20, 30, 50, 200, 300, 400, and 500. Topic coherence was calculated using *the Gensim CoherenceModel* with the C_v measure. The C_v coherence score evaluates topic interpretability based on the co-occurrence and semantic similarity of representative topic words. The overall score is obtained by aggregating the coherence values across all topics.

2.4. Hyperparameter tuning with combinatorial optimization approach (grid search)

Hyperparameter tuning in LDA is the process of selecting model settings by varying three main parameters: k, α, and β. The parameter k determines how many topics are identified, α controls how concentrated a document's topic distribution is, and β controls how concentrated the word distribution within each topic is. This study uses the C_v coherence score as the evaluation metric. The LDA model is trained repeatedly across different

combinations of *k*, *α*, and *β* on both 75% and 100% corpus configurations. Because coherence is calculated with respect to the corpus used for each model, scores from the two corpus sizes are not treated as directly comparable. The final configuration is therefore selected from the models trained on the complete 100% corpus, with the highest C_v coherence among those configurations used for the final analysis.

a. Define the evaluation function. A supporting function called *compute_coherence_values* is created, which trains the LDA model with certain hyperparameter combinations, then calculates the coherence value (C_v) as an evaluation metric. The C_v coherence score measures the extent to which words in each topic co-occur within documents, so that the resulting topic becomes more interpretable.
b. Define the hyperparameter grid. Specify the range of values to be tested for each hyperparameter, namely: *k*, e.g., from 1 to 20; *α*: numerical values such as 0.1, 0.4, 0.7, as well as 'symmetric' and 'asymmetric' options, and *β*: numerical values such as 0.1, 0.4, 0.7, as well as a 'symmetric' option.
c. Perform grid search and experimentation, combinatorial exploration is carried out on all possible combinations of hyperparameters, by: training LDA models on a subset of data (e.g., 75% of the corpus) and the entire data (100% of the corpus); iterating over each combination (*k*, *α*, *β*); calculating and recording the C_v coherence score for each model, and storing the evaluation results of each combination in a data structure (*dictionary*) and exporting them to a CSV file for further analysis.
d. Select the optimal hyperparameters, the hyperparameter combination with the highest coherence score is selected as the final configuration to train the LDA model used in the subsequent analysis stages, such as intertopic distance map visualization.

The combinatorial grid search explored 20 candidate values of *k* (1–20), 5 values of *α* (0.1, 0.4, 0.7, symmetric, asymmetric), and 4 values of *β* (0.1, 0.4, 0.7, symmetric) across two corpus subsets (75% and 100%), resulting in 20×5×4×2=800 \times 5 \times 4 \times 2=800 LDA configurations. Each configuration used a fixed random_state (Table 1) to minimize stochastic variation, and the search was terminated after exhaustive evaluation of the predefined grid. In this study, "*combinatorial optimization*" specifically refers to exhaustive grid search over discretized hyperparameter combinations, rather than random search, Bayesian optimization, or genetic search. Direct comparison with these alternative strategies was beyond the scope of this study and remains a direction for future research.

3. RESULTS AND DISCUSSION

In the hyperparameter tuning stage of the LDA model, a combinatorial exploration of various hyperparameter configurations was carried out to obtain the model with the best topic coherence. This tuning process was carried out through a grid search that included variations in the *k*, *α*, and *β*, each tested on two data subsets (75% corpus and 100% corpus). For each hyperparameter combination, the LDA model was trained and evaluated using the C_v coherence score. Based on the evaluation results, an optimal hyperparameter configuration was obtained that produced the highest coherence score, which was subsequently used as the final parameter configuration in the topic modeling process.

3.1. Phrase modeling results

The phrase modeling stage includes stopword removal, bigram and trigram construction, and lemmatization. The resulting text consists of normalized words and phrase tokens that are used for subsequent LDA modeling. For example, lemmatization converts "*sensors*" to "*sensor*" and "*studied*" to "*study*", while frequently co-occurring terms are combined into phrase tokens such as "*chlorophyll_fluorescence*", "*shoot_length*", "*positively_correlate*", and "*negatively_correlate*". These phrase tokens preserve contextual relationships between words and provide a more consistent representation for topic modeling.

3.2. Data transformation results

In LDA topic modeling, the main inputs are the dictionary (*id2word*) and the corpus. The dictionary assigns a unique integer ID to each word in the lemmatized and phrase-modeled data, which is then used to construct the corpus in BoW format. Each document is represented as a list of *(word ID, frequency)* tuples. For example, *(*0, 1) indicates that word ID 0 occurs once, while (2, 5) indicates five occurrences of word ID 2. This numerical representation converts the textual corpus into a format that can be processed by LDA for topic distribution analysis.

3.3. Trial results of number of topics and chunk size

In the trial stage, the base model was built using various combinations of *k* and chunk size to identify the initial configuration that produced the highest coherence value, calculated using the C_v coherence score as the evaluation metric. The test results for *k* 5 and 10 are shown in Table 3, while the results for *k* 15 and 20 are shown in Table 4. Based on these results, the model with 10 topics achieved a coherence value of 0.5063 at a chunk size of 300, the model with 15 topics achieved a coherence value of 0.5692 at a chunk size of 200, and

the model with 20 topics achieved the highest coherence value of 0.6029 at a chunk size of 10. However, considering processing efficiency, the second-highest coherence value for the 20-topic model, 0.5737 at a chunk size of 200, was selected as the parameter configuration for training the base LDA model.

Table 3. Trial *k* and chunk sizes (5 and 10)

5 topics		10 topics	
Chunk size	Coherence score	Chunk size	Coherence score
10	0.42607966921	10	0.48897544298
20	0.41794738883	20	0.48437950131
30	0.43331177080	30	0.45304384388
50	0.41317654163	50	0.49768116693
200	0.42496855067	200	0.44696391787
300	0.43976090237	300	**0.50639179455**
400	0.45806567913	400	0.46250318463
500	0.41410387166	500	0.41288173999

Table 4. Trial *k* and chunk sizes (15 and 20)

15 topics		20 topics	
Chunk size	Coherence score	Chunk size	Coherence score
10	**0.51975470851**	10	**0.60299465571**
20	0.46572219964	20	**0.53620547883**
30	0.47612058904	30	0.48565341644
50	0.40974341707	50	**0.51621368550**
200	**0.56927150526**	200	**0.57379130994**
300	0.47350007813	300	**0.51943278791**
400	0.48282817461	400	**0.51043568617**
500	0.42050121104	500	**0.50047000225**

3.4. Results of word weight calculation on topics

At this stage, the LDA model was trained using the optimal number of topics and chunk size obtained from the previous trial stage. The output of the model is a list of keywords for each topic along with their weights, representing the relative contribution of each word to the formation of that topic. For example, in Topic 1, the word "*soil*" has the highest weight (0.027), followed by "*water*" (0.014), "*crop*" (0.010), and "*yield*" (0.009). In contrast, several other topics, such as Topic 2, Topic 3, and Topic 5, contain words with very low or near-zero weights, indicating that these topics were not well formed. The complete results of the word weight calculation are presented in Table 5, showing the variation in word contribution across the topics generated by the model.

Table 5. Top keywords and weights for selected topics

Topic	Top keywords (weight)
Topic 1	soil (0.027), water (0.014), crop (0.010), increase (0.010), study (0.010), yield (0.009), high (0.008), area (0.008), result (0.007), use (0.007).
Topic 2	peten (0.000), campechano (0.000), preclassic (0.000), mimosin (0.000), eluate (0.000), tfff (0.000), basamid (0.000), giant_leucaena (0.000), kaan (0.000), seventh_century (0.000).
Topic 3	futurity (0.000), dravidian (0.000), circense (0.000), artist (0.000), irl (0.000), shudra (0.000), dalit (0.000), utopia (0.000), landmass (0.000), dly (0.000).
Topic 20	olive (0.016), olive_grove (0.001), humanlabele (0.000), photorealistic (0.000), autolabele (0.000), ascospore (0.000), olivegrowe (0.000), evoo (0.000), ammonium_bicarbonate (0.000), midlate (0.000).

Based on Table 5, the topics can be distinguished according to their agricultural relevance and coherence. Topic 1, with keywords such as "*soil*", "*water*", "*crop*" and "*yield*" represents crop productivity and environmental factors. Topic 4 focuses on livestock nutrition and management through terms such as "*animal*", "*feed*", "*diet*" and "*cattle*" while Topics 12 and 14 relate to plant research and genetics. Topic 7 highlights environmental issues, including carbon emissions and biochar. In contrast, Topics 2, 3, and 5 contain seemingly unrelated or near-zero-weight terms, indicating noise or rare vocabulary. Such topics may occur when the specified number of topics is relatively large compared with the separable themes in the corpus and are commonly referred to as "*garbage topics*".

A closer examination of Topics 2 and 3 shows terms such as "*Peten*", "*Campechano*", "*Preclassic*", "*Dravidian*", "*Dalit*", "*mimosine*", and "*eluate*" reflecting proper nouns, foreign-language fragments, and highly specific scientific terms with limited co-occurrence. This variation is partly attributable to the broad use of the keyword "*Agriculture*" across three heterogeneous databases, which may retrieve publications in which agricultural terms are incidental to other domains. Consequently, LDA may allocate some topics to rare, low-

frequency terms, producing the near-zero weights observed in Table 4. Overall, word weights help assess keyword contributions and support subsequent topic interpretation, visualization, and model validation.

3.5. Hyperparameter tuning results

A combinatorial grid search was conducted to determine the optimal configuration for the final LDA model by varying k (1–20), α (0.1, 0.4, 0.7, symmetric, and asymmetric), and β (0.1, 0.4, 0.7, and symmetric) across 75% and 100% corpus subsets. The highest-scoring configurations are summarized in Table 6.

Table 6. Hyperparameter tuning results (k, α, and β)

Corpus (%)	Topics	α	β	Coherence
75	15	0.1	0.70	0.5364
75	9	0.1	0.70	0.5332
100	**17**	**0.1**	**0.70**	**0.5075**

The highest coherence score on the 75% corpus was 0.5364 with 15 topics, α=0.1 and β=0.70. However, this model was trained on only part of the corpus. For the complete 100% corpus, the highest coherence score was 0.5075 with 17 topics, α=0.1, and β=0.70. This configuration was selected for the final LDA model and subsequent topic visualization.

To assess whether the upper bound of k=20 and the selection of 17 topics were arbitrary, Table 7 summarizes the coherence trend across the complete range of k=1–20 for α=0.1 on the full (100%) corpus, for the two β settings for which complete results across this range were retained from the original grid-search output (β=0.70 and β=symmetric).

As shown in Table 7, coherence increases from k=1 and reaches its highest value at k=17 (0.5075) for β=0.70, followed by a decline at k=18–20. Thus, the upper bound of k=20 was sufficient to capture the peak rather than truncate an increasing trend. The values for k=8–19 show a mean coherence of approximately 0.465 for β=0.70 and 0.461 for β=symmetric, indicating that k=17 lies within a relatively strong region of the search space rather than representing an isolated peak. However, this analysis does not establish topic stability across different random seeds, which was not evaluated in this study.

Table 7. Coherence score trend across number of topics (k=1–20, α=0.1, and 100% corpus)

k	Coherence (β=0.70)	Coherence (β=symmetric)
1	0.3017	0.2906
2	0.4013	0.4013
3	0.3977	0.4025
4	0.4082	0.4025
5	0.4233	0.4207
6	0.4440	0.4039
7	0.4293	0.4426
8	0.4809	0.4389
9	0.4699	0.4468
10	0.4664	0.4492
11	0.4863	0.4607
12	0.4733	0.4683
13	0.4629	0.4557
14	0.4557	0.4631
15	0.4748	0.4684
16	0.4569	0.4728
17*	**0.5075**	**0.4751**
18	0.4305	0.4643
19	0.4108	0.4704
20	0.4350	0.4611

* Selected final-model configuration (k=17, α=0.1, β=0.70, and 100% corpus)

Figure 2 illustrates the coherence score trend across different k for the 100% corpus with α=0.1. The figure compares two β configurations, namely β=0.70 and symmetric β. The results show that the coherence scores generally increase as the number of topics increases, although some fluctuations are observed. The highest coherence score is obtained at k=17, reaching approximately 0.51 for β=0.70, which indicates that 17 topics provide the most coherent topic structure among the evaluated configurations. Therefore, k=17 was selected as the optimal number of topics for subsequent LDA modeling.

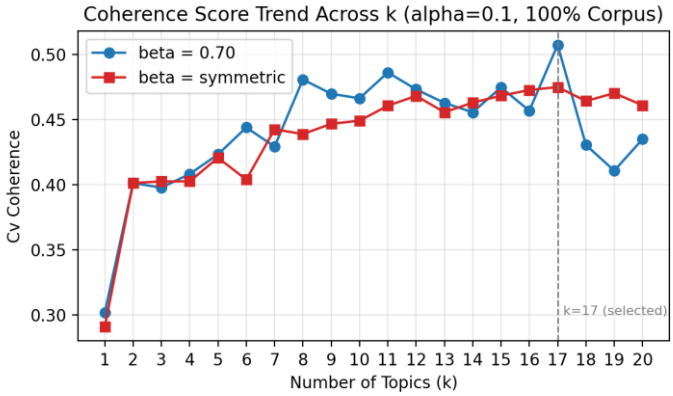


Figure 2. Coherence score trend across k (α=0.1, 100% corpus)

Regarding the discrepancy between the two highest-scoring configurations on the 75% corpus (15 topics, C_v=0.5364; 9 topics, C_v=0.5332) and the selected 100% corpus configuration (17 topics, C_v=0.5075), C_v coherence is computed using the corpus on which each model is trained. Therefore, coherence values from the 75% and 100% corpora are not directly comparable because they use different reference corpora and vocabularies. The 100% configuration was selected as the highest-scoring model evaluated on the complete dataset, ensuring that all publications contribute to the final topic structure and visualization. This selection reflects dataset representativeness rather than formal independent hold-out validation, which was not performed. Such validation is recommended for future studies.

3.6. Intertopic distance map visualization results

The final LDA model was retrained using α=0.1, β=0.7, and 17 topics and visualized using the intertopic distance map, as shown in Figure 3. Figure 3 shows the spatial distribution of the 17 topics, where circle size indicates topic prevalence and spatial proximity reflect similarity in topic-word distributions. Topic 1 is the most prevalent and overlaps with Topics 2 and 4, while Topics 3, 5, 6, and 8 form another closely related group. Topic 7 appears relatively isolated, whereas several smaller topics cluster near the origin, indicating lower prevalence and potential thematic redundancy. These relationships provide a visual basis for interpreting the topic structure alongside the keywords in Table 4.

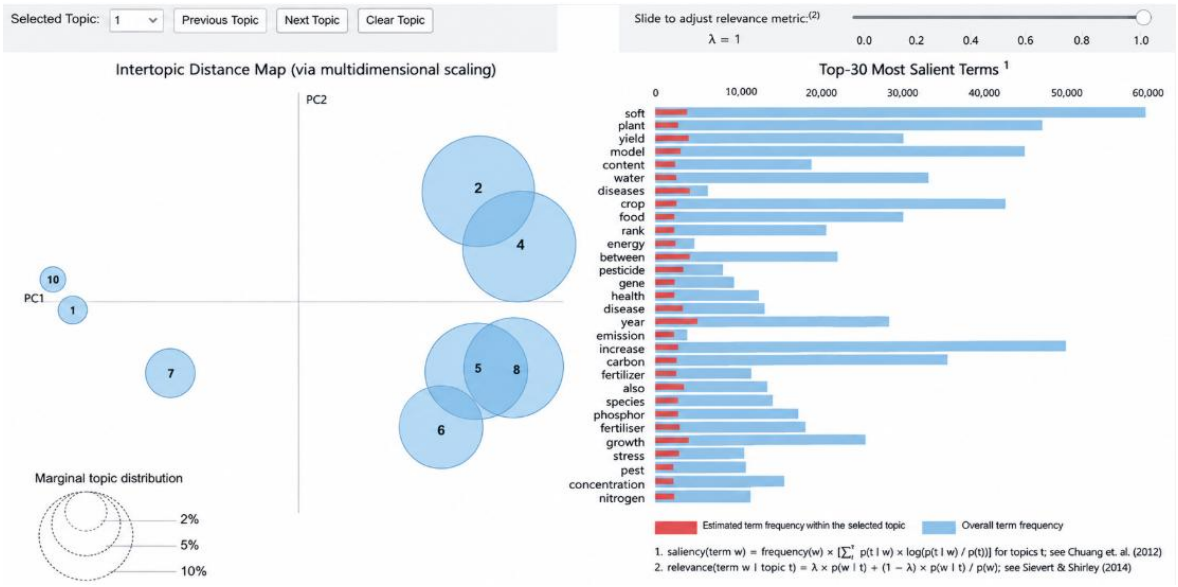


Figure 3. LDA model visualization with intertopic distance map

The saliency panel shows dominant terms including soil, plant, yield, model, control, water, carbon, crop, and food, reflecting major themes related to agricultural productivity, environmental factors, and management practices. Thus, the visualization complements the word-weight and coherence analyses. Methodologically, *pyLDAvis* represents topic-word distributions using pairwise Jensen–Shannon divergence and projects them into two dimensions using multidimensional scaling. Circle size represents topic prevalence in the corpus. Because the map is a two-dimensional projection, topic overlap is interpreted as a visual indication of similarity rather than conclusive evidence of redundancy. Quantitative topic-similarity or diversity measures were not computed in this study.

To illustrate topic-level interpretation, Figure 4 presents Topic 1 as a representative example. Topic 1 has the largest circle, representing approximately 26% of corpus tokens, and is characterized by keywords such as agricultural, food, farmer, agriculture, production, system, and development. These terms indicate a broad theme related to agricultural practices, food systems, and production. Its overlap with Topics 2 and 4 suggests shared vocabulary, while its separation from Topics 7 and 10 indicates greater thematic distinction.

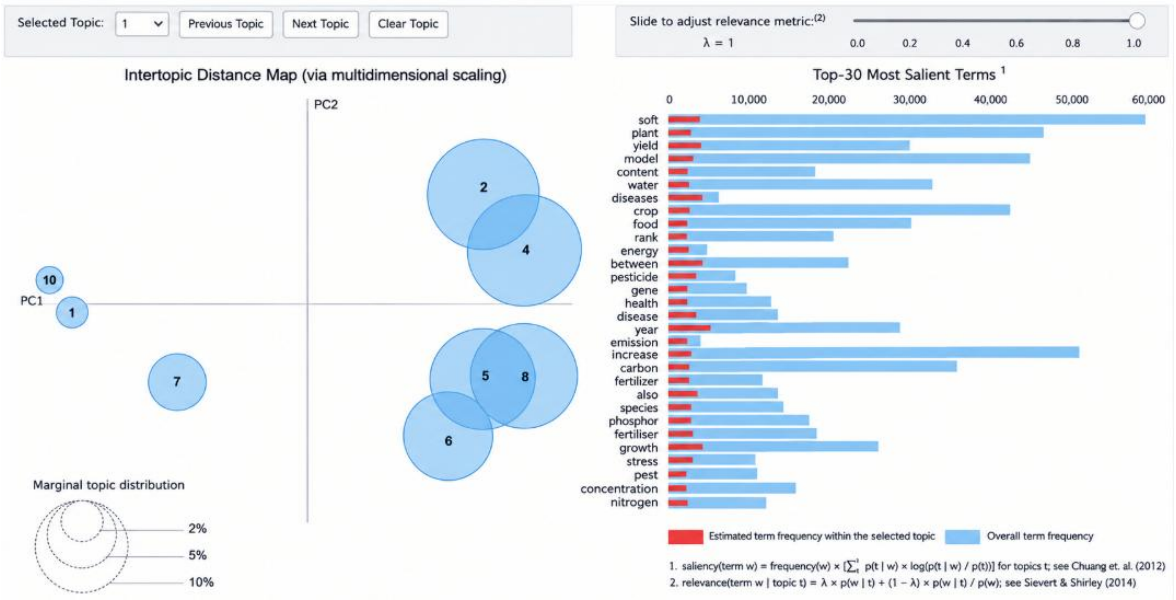


Figure 4. Visualization of the LDA model for Topic 1

The other major topics are summarized as follows. Topic 2 (17.9%) focuses on computational models and data-analysis methods; Topic 3 (16%) represents plant production and growth; Topic 4 (15.2%) concerns environmental and land-use change; Topic 5 (11.2%) reflects experimental and sample-based analysis; and Topic 6 (8.7%) focuses on plant genetics and environmental stress. Together, Topics 1–6 account for approximately 95% of corpus tokens, while the remaining topics form smaller clusters with greater overlap and, in some cases, near-zero weights.

From an information-system perspective, the visualization supports interpretation of topic prevalence and relationships, which may facilitate scholarly search, document organization, and knowledge discovery. These represent potential applications rather than functions implemented or evaluated in this study. The observed residual noise may be reduced in future work through document-frequency filtering, domain-specific stopwords, and more specific retrieval queries.

Overall, the results demonstrate that the optimized LDA model and intertopic distance map provide complementary quantitative and visual information for interpreting agricultural research topics. However, without comparison with an untuned baseline or alternative topic-modeling methods, the selected configuration cannot be considered universally superior. The six thematic groups identified here are specific to the analyzed corpus and should not be interpreted as definitive trends in agricultural research.

3.7. Study limitations and reproducibility considerations

Several limitations should be considered when interpreting the results. The evaluation relied mainly on C_v coherence and qualitative topic inspection, as topic diversity, redundancy, and stability across random seeds were not evaluated. A formal ablation study was also not conducted to quantify the individual effects of

preprocessing components, corpus size, and LDA hyperparameters. In addition, no comparison was made with untuned LDA or alternative topic-modeling methods such as NMF, LSA, or BERTopic; therefore, the reported improvement applies only to the evaluated LDA configurations. Detailed records of runtime, memory usage, hardware, source-level deduplication, and exact software versions were also not systematically retained, limiting complete computational reproducibility. These limitations do not invalidate the reported results but indicate that future studies should include comprehensive topic-quality evaluation, component-level ablation, alternative-model comparison, and systematic recording of data-processing and computational configurations.

4. CONCLUSION

This study applied a systematic, grid-search-based hyperparameter optimization approach to determine LDA parameters for modeling agricultural research topics from large-scale scientific publication data covering 2012–2023. Its methodological contribution lies in combining exhaustive search over LDA parameter combinations with intertopic distance map visualization to support interpretation of the resulting topic structure. From the 75,031 publications analyzed, the final model used 17 topics with α=0.1 and β=0.70, corresponding to the highest C_v coherence among the configurations trained on the complete 100% corpus. Six dominant thematic groups were identified: agricultural practices and food systems, data analysis applications, plant production and growth factors, environmental and land-use changes, effects and sample analysis, and plant genetics under environmental pressures. Because the model aggregates publications across the entire 2012–2023 period, the results represent a static thematic structure rather than a direct measurement of temporal research trends. Likewise, the study does not establish superiority over alternative topic-modeling methods because no such comparative experiment was conducted. Future work should evaluate topic diversity and stability across random seeds, compare LDA with alternative approaches, and apply dynamic topic modeling and citation analysis to examine topic evolution. The workflow can potentially be adapted to other large, multi-source scholarly corpora, subject to domain-specific preprocessing and hyperparameter retuning.

ACKNOWLEDGMENTS

The authors would like to express their sincere appreciation to the Research Institutions of Gunadarma University for their support and facilitation throughout this research.

FUNDING INFORMATION

This research did not receive any external research grant or contract.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Maukar	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Diana Tri Susetianingtias		✓	✓			✓			✓					
Eka Patriya						✓	✓	✓	✓					
Rodiah			✓			✓			✓		✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The preprocessing scripts, the hyperparameter-search configuration, and the aggregated coherence result underlying Tables 6 and 7 are available from the corresponding author upon reasonable request. The complete raw bibliographic dataset cannot be publicly shared because the source databases are subscription-based and subject to their access and use conditions.

REFERENCES

[1] A. Gamage *et al.*, "Advancing sustainability: The impact of emerging technologies in agriculture," *Current Plant Biology*, vol. 40, p. 100420, Dec. 2024, doi: 10.1016/j.cpb.2024.100420.

[2] L. N. Yogi, T. Thalal, and S. Bhandari, "The role of agriculture in Nepal's economic development: Challenges, opportunities, and pathways for modernization," *Heliyon*, vol. 11, no. 2, p. e41860, Jan. 2025, doi: 10.1016/j.heliyon.2025.e41860.

[3] G. Mgendi, "Unlocking the potential of precision agriculture for sustainable farming," *Discover Agriculture*, vol. 2, no. 1, p. 87, Nov. 2024, doi: 10.1007/s44279-024-00078-3.

[4] N. Aijaz, H. Lan, T. Raza, M. Yaqub, R. Iqbal, and M. S. Pathan, "Artificial intelligence in agriculture: Advancing crop productivity and sustainability," *Journal of Agriculture and Food Research*, vol. 20, p. 101762, Apr. 2025, doi: 10.1016/j.jafr.2025.101762.

[5] R. Espinel, G. Herrera-Franco, J. L. R. García, and P. Escandón-Panchana, "Artificial Intelligence in Agricultural Mapping: A Review," *Agriculture (Switzerland)*, vol. 14, no. 7, p. 1071, Jul. 2024, doi: 10.3390/agriculture14071071.

[6] A. M. Grisales, S. Robledo, and M. Zuluaga, "Topic Modeling: Perspectives from a Literature Review," *IEEE Access*, vol. 11, pp. 4066–4078, 2023, doi: 10.1109/ACCESS.2022.3232939.

[7] U. C. Calistus, M. O. Onyesolu, A. C. Doris, and C. V. Egwu, "Exploring Latent Dirichlet Allocation (LDA) in Topic Modeling: Theory, Applications, and Future Directions," *Newport International Journal of Engineering and Physical Sciences*, vol. 4, no. 1, pp. 9–16, Mar. 2024, doi: 10.59298/nijep/2024/41916.1.1100.

[8] S. Chiangnangam, P. Manorom, and W. Chansanam, "Analyzing Research Trends in Organic Agriculture and Entrepreneurship: A Topic Modeling Approach Using Latent Dirichlet Allocation (LDA)," *Journal of Computational Analysis and Applications (JoCAAA)*, vol. 33, no. 5, pp. 191–203, 2024.

[9] L. T. Nguyen *et al.*, "Evaluating the Performance of Topic Modeling Techniques for Bibliometric Analysis Research: An LDA-based Approach," *HighTech and Innovation Journal*, vol. 5, no. 2, pp. 312–330, Jun. 2024, doi: 10.28991/HIJ-2024-05-02-07.

[10] P. C. Kallens and M. H. Christiansen, "Models of Language and Multiword Expressions," *Frontiers in Artificial Intelligence*, vol. 5, 2022, doi: 10.3389/frai.2022.781962.

[11] R. Mao, K. He, C. B. Ong, Q. Liu, and E. Cambria, "MetaPro 2.0: Computational Metaphor Processing on the Effectiveness of Anomalous Language Modeling," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 9891–9908, doi: 10.18653/v1/2024.findings-acl.590.

[12] B. Ogunleye, B. S. L. Barrantes, and K. I. Zakariyyah, "Topic modelling through the bibliometrics lens and its technique," *Artificial Intelligence Review*, vol. 58, no. 3, p. 74, Jan. 2025, doi: 10.1007/s10462-024-11011-x.

[13] Y. Setiawan, N. U. Maulidevi, and K. Surendro, "The Optimization of n-Gram Feature Extraction Based on Term Occurrence for Cyberbullying Classification," *Data Science Journal*, vol. 23, no. 1, May 2024, doi: 10.5334/dsj-2024-031.

[14] N. Nasser, L. Karim, A. E. Ouadrhiri, A. Ali, and N. Khan, "n-Gram based language processing using Twitter dataset to identify COVID-19 patients," *Sustainable Cities and Society*, vol. 72, p. 103048, Sep. 2021, doi: 10.1016/j.scs.2021.103048.

[15] Y. Guntzburger and M. Hadengue, "Providing a comprehensive thematic review of the Science-Policy Interface (SPI): A probabilistic topic modeling approach," *Environmental Science and Policy*, vol. 163, p. 103966, Jan. 2025, doi: 10.1016/j.envsci.2024.103966.

[16] O. Ozyurt, H. Özköse, and A. Ayaz, "Evaluating the latest trends of Industry 4.0 based on LDA topic model," *Journal of Supercomputing*, vol. 80, no. 13, pp. 19003–19030, Sep. 2024, doi: 10.1007/s11227-024-06247-x.

[17] S. Sarica and J. Luo, "Stopwords in technical language processing," *PLoS ONE*, vol. 16, no. 8, p. e0254937, Aug. 2021, doi: 10.1371/journal.pone.0254937.

[18] A. Mor, C. Orsenigo, M. S. Gomez, and C. Vercellis, "Shaping the causes of product returns: topic modeling on online customer reviews," *Electronic Commerce Research*, vol. 26, no. 1, pp. 747–781, Feb. 2026, doi: 10.1007/s10660-024-09901-x.

[19] M. Almgerbi, A. De Mauro, A. Kahlawi, and V. Poggioni, "Improving Topic Modeling Performance through N-gram Removal," in *ACM International Conference Proceeding Series*, Dec. 2021, pp. 162–169, doi: 10.1145/3486622.3493952.

[20] K. E. Chu, P. Keikhosrokiani, and M. P. Asl, "A Topic Modeling and Sentiment Analysis Model for Detection and Visualization of Themes in Literary Texts," *Pertanika Journal of Science and Technology*, vol. 30, no. 4, pp. 2535–2561, Sep. 2022, doi: 10.47836/pjst.30.4.14.

[21] E. Cambria, "Syntactics Processing," in *Understanding Natural Language Understanding*, Cham: Springer Nature Switzerland, 2025, pp. 41–112, doi: 10.1007/978-3-031-73974-3_2.

[22] S. Chen and N. R. M. Nordin, "Effectiveness of Morphological Intervention on Two Measures of Vocabulary Knowledge and Listening Comprehension," *World Journal of English Language*, vol. 14, no. 4, pp. 336–349, Apr. 2024, doi: 10.5430/wjel.v14n4p336.

[23] D. Angelov and D. Inkpen, "Topic Modeling: Contextual Token Embeddings Are All You Need," in *EMNLP 2024-2024 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP 2024*, 2024, pp. 13528–13539, doi: 10.18653/v1/2024.findings-emnlp.790.

[24] A. Farea, S. Tripathi, G. Glazko, and F. Emmert-Streib, "Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108877, Oct. 2024, doi: 10.1016/j.engappai.2024.108877.

[25] A. Giaconia, V. Chiariello, S. Giannuzzi, and M. Passarotti, "Topic Modeling for Auditing Purposes in the Banking Sector," in *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, Dec. 2024, vol. 3878, pp. 1030–1035.

[26] O. Ozyurt and A. Ayaz, "Twenty-five years of education and information technologies: Insights from a topic modeling based bibliometric analysis," *Education and Information Technologies*, vol. 27, no. 8, pp. 11025–11054, Sep. 2022, doi: 10.1007/s10639-022-11071-y.

[27] J. C. Campbell, A. Hindle, and E. Stroulia, "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," in *The Art and Science of Analyzing Software Data*, Elsevier, 2015, pp. 139–159, doi: 10.1016/B978-0-12-411519-4.00006-9.

[28] N. M. N. Mathivanan, R. M. Janor, S. A. Razak, and N. A. M. Ghani, "Feature Substitution Using Latent Dirichlet Allocation for Text Classification," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, pp. 1087–1098, 2025, doi: 10.14569/IJACSA.2025.01601105.
[29] H. Özköse, O. Ozyurt, and A. Ayaz, "Management Information Systems Research: A Topic Modeling Based Bibliometric Analysis," *Journal of Computer Information Systems*, vol. 63, no. 5, pp. 1166–1182, Sep. 2023, doi: 10.1080/08874417.2022.2132429.
[30] S. Bellaouar, A. Itbirene, and B. Chihani, "Word Embedding-based Topic Modeling," in *Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)*, Amsterdam, Netherlands: Atlantis Press International BV, 2024, doi: 10.2991/978-94-6463-496-9_8.

BIOGRAPHIES OF AUTHORS



Maukar is a leading professional in the field of Computer Science Armed with a bachelor's degree in Computer Science from the University of Indonesia which was completed in 1991, He is continued his education by earning a master's degree in Technology Management from the Bandung Institute of Technology in 1994. His academic dedication culminated in obtaining a doctoral degree in Information Technology from Gunadarma University in 2014. He can be contacted at email: maukar@staff.gunadarma.ac.id.



Diana Tri Susetianingtias is currently researcher and lecturer at Gunadarma University. From 2017 until now, won 4 research grants from Indonesian Directorate General for Higher Education DIKTI (RISTEKDIKTI). Nowadays, authoring 2 books about medical image processing for retinal fundus image, and 1 book about retinal biometric. She can be contacted at email: diants@staff.gunadarma.ac.id.



Eka Patriya is bachelor's degree in System Informatic from Gunadarma University in 1997 and he get master's degree in Finance Management from the University of Indonesia in 2004. He is currently a lecturer at the Faculty of Management, Gunadarma University, Jakarta, Indonesia. He can be contacted at email: ekapatriya@staff.gunadarma.ac.id.



Rodiah is an associate professor, researcher, and lecturer, and currently serves as the vice head of Postgraduate Academic System Development at Gunadarma University. Since 2012, she has received nine competitive research grants from the Indonesian Directorate General of Higher Education (DIKTI, Ministry of Research and Technology). Her research interests primarily focus on medical image processing, with particular emphasis on retinal fundus imaging. She has authored two books on medical image processing for retinal fundus images and one book on retinal biometrics. She can be contacted at email: rodiah@staff.gunadarma.ac.id.