J. Intell. Comput. Health Inform. Vol. 3, No. 2, 2022
ISSN: 2715-6923 e-ISSN: 2721-9186, pp.35~46 | DOI: 10.26714/jichi.v3i2.11428

Topic Modelling using Latent Dirichlet Allocation (LDA) to Investigate the Latent Topics of Mathematical Creative Thinking Research in Indonesia

Della Maulidiya

Mathematics Education, Teacher Training and Education Faculty, Bengkulu University, Jl. WR Supratman Kandang Limun Bengkulu, Indonesia
*Corresponding author: della.maulidiya@unib.ac.id

Article histroy :
Received: 14 SEP 2022
Accepted: 26 JAN 2023
Avalaible online: 23 FEB 2023

Research article

Abstract: In mathematics education, there is increasing interest in Mathematical Creative Thinking (MCT), and numerous scientific documents on this topic have been published in Indonesia. The availability of publication databases has made it easier to access these documents and collect large amounts of data related to MCT. This data has the potential to uncover latent topics within MCT research articles published in Indonesian journals. This study analyzed a dataset of 102 articles obtained from Garuda (Digital Reference Garda) published between 2010 and 2022 in six proceedings and 49 journals. The study applied text processing techniques and used topic modeling with Latent Dirichlet Allocation (LDA) and variational expectation maximization algorithm (VEM) to produce 23 topics. Each topic consisted of general and special words from the beta probability value. The study found 30 unique words from topic modeling, including learning, abilities, problems, skills, mathematics, tests, levels, answers, approaches, assessments, basics, classes, developing, values, ideas, instruments, materials, mathematics, moderate, models, motivation, open-ended, processes, questions, reasons, solving, styles, subjects, teaching, and worksheets. The study also used LDA to classify documents into discovered themes and found that the five MCT research focuses were learning approaches, student competencies, teacher competencies, assessments, and learning resources. The study's findings revealed a research gap, specifically, the need for more MCT studies that concentrate on enhancing teacher competency.

Keywords: TOPIC MODELLING; LATENT DIRICHLET ALLOCATION; MATHEMATICAL CREATIVE THINKING; LATENT TOPIC; TEXT PROCESSING

1. Introduction

The application of machine learning algorithms in education has been carried out for various needs. For example, (Syaputri, et al., 2020) use the Naive Bayes algorithm to classify majors based on the average grade of high school students for three semesters. On the other hand, machine learning is also applied to text processing, as done by (Hwang & Cho, 2021), to identify research topics and trends related to math teacher knowledge based on scientific articles. Their used topic modeling with the Latent Dirichlet Allocation (LDA) algorithm to overcome the limitations of manual coding methods in analyzing and categorizing broad text data. The two research examples show that the availability of big data, both numeric and text data, in the field of education can be done with the help of machine learning.

Search results on Google Scholar using the keyword "mathematical creative thinking" returns more than 15,000 publications. These results indicate the great interest of researchers in educational mathematics in this study. Mathematical creative thinking (MCT) is essential for students to process or realize products (Suherman & Vidákovich, 2022). The Program for International Student Assessment (PISA) defines MCT as the ability to actively engage in learning, assessing, and improving concepts that can lead to valuable new solutions (OECD, 2019). When solving problems, the ability to develop original ideas or solutions is called creative thinking (Hadar & Tirosh, 2019).

With the demand for future generations to be able to innovate and be creative, the ability to think creatively is one of the learning outcomes goals, including mathematics. In mathematics, creative behavior results from the ability to solve and model problems, build new concepts, and recognize new relationships that add to the body of knowledge, raise new questions, or reframe existing ones (Newton et al., 2022). If the work is more complicated, it could be harder to generate useful uniqueness through divergent thought (de Vink et al., 2022; Siswono, 2010). It is critical for educators to be aware of individual differences in creative thinking and their potential impact on mathematics ability, according to the interaction effects discovered by (de Vink et al., 2022). Therefore, further research is required to determine the relationship between creative thinking and mathematical ability and strategies for fostering it. This phenomenon encourages various efforts to improve the quality of learning and to teach mathematics to improve MCT skills which are then published as scientific articles.

Numerous initiatives to promote MCT have been described in Indonesia's hundreds of MCT research articles. A literature review is necessary to explore the MCT research to understand, organize, and label this text data. However, a manual literature review requires time and effort. The manual method tends to be inaccurate because it involves a tiring and tedious process. Therefore it is necessary to use computational methods suitable for large corpus (Chen et al., 2020). In natural language processing, topic modeling techniques can be very helpful and productive for semantic mining and latent finding in datasets of documents (Jelodar et al., 2019). For research topic detection and tracking, topic modeling is considered more flexible and effective because it automatically finds semantic topics from the text corpus without training (Gurcan, 2018; Hwang & Cho, 2021; Vayansky & Kumar, 2020). Topic modeling has proven to be a valuable and suitable tool for uncovering meaningful topics from large amounts of text data (Jelodar et al., 2019).

One of the most widely used topic modeling algorithms is LDA which is considered flexible and adaptive (Jelodar et al., 2019; Vayansky & Kumar, 2020). Most of the newly published research on topic modeling methods derives from LDA or combines it with other models for many purposes, making LDA the most widely cited method in topic modeling research (Blei, 2012; Jelodar et al., 2019; Vayansky & Kumar, 2020). Although still limited, some recent studies have applied topic modeling using LDA to educational research (e.g., (Chen et al., 2020; Hwang & Cho, 2021)). Ones can be criticized from various LDA topic modeling articles is the lack of detailed explanation regarding the stages of its application, from determining the source, searching, cleaning, and preprocessing to data analysis. This article systematically presents these stages based on the process by (Hwang & Cho, 2021) and (Gürcan, 2018).

In LDA, documents are assumed to be generated by randomly combining hidden topics, which are seen as a probability distribution of words (Blei, 2012; Vayansky & Kumar, 2020). However, one of the problems with using journals as a source of topic modeling data is that there are sections that unrelevant to the research topics, such as acknowledgments and references. In addition, journals published in Indonesia generally provide bilingual abstracts (English and Indonesian). (Denny & Spirling, 2018) mention that one of the crucial points of using text processing with unsupervised learning is preprocessing. Therefore, this research demonstrates how the preprocessing stage is carried out to prepare the corpus for topic modeling.

On the other hand, most topic modeling research uses the Scopus database, Web of Science (WoS), or journals outside Indonesia. The results of observations on Google Scholar show that many English-language journals are published in Indonesia but have yet to be indexed by Scopus or Wos. Indonesia's journal articles can describe hidden topics of Indonesian research more effectively if they are investigated in depth by text processing methods, such as LDA topic modelling. Therefore, this study did not focus on LDA optimization but applied the LDA algorithm for topic modeling to explore hidden (latent) topics in MCT publications from a database of Indonesian scientific publications.

2. Methods

Topic modeling is a method of computational and statistical modeling analysis in text processing where many abstract topics are found in a corpus (a massive collection of documents). Topics are statistically derived and performed automatically based on words that frequently occur together in the corpus (Hwang & Cho, 2021). Each document is considered in topic modeling as a mixture of themes, each topic as a distribution of terms, and each document as a collection of terms (Chen et al., 2020). In (Blei, 2012) stated that thematic structure would be a new window for exploring and digesting unstructured document sets. As an "unsupervised" machine learning technique, topic modeling requires no prior training (Gürcan, 2018), in contrast to topic classification systems, which need to be trained before being able to analyze text automatically. Therefore, there is no preset description of the subjects that will emerge from the text and how the topics are related to words and documents before the analysis is done.

This study choose LDA as a topic modeling algorithm because it can find hidden (latent) semantic patterns from documents in the big data literature (Gürcan, 2018). The LDA algorithm was published in 2003 by David M. Blei, Andrew Y. Ng, and Michael I. Jordan. LDA is a generative three-level hierarchical Bayesian probabilistic model for discrete data sets. In generative probabilistic modeling, data is treated as the output of a generative process that includes hidden variables (Blei, 2012). He explained that the generative process defines the combined probability distribution of observed and hidden random variables.

The hidden structure in LDA is called a topic, where each topic is defined as a distribution of words, and each document is a combined distribution of topics (Gürcan, 2018). Formally the generative process for LDA is expressed by the following notation. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary. The topic proportions for the *d*-th document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic *k* in document *d*. The topic assignments for the *d*-th document are z_d, where $z_{d,n}$ is the topic assignment for the *n*-th word in document *d*. The observed words for document *d* are w_d, where $w_{d,n}$ is the *n*-th word in *d*.

The assumptions made by LDA are 1) a "bag of words," namely, the order of words in a document is not essential, 2) the order of documents has no effect, and 3) the number of topics is assumed to be known and fixed (Blei, 2012). These three assumptions are the key points of the topic modeling process using LDA and can be illustrated in Fig. 1.

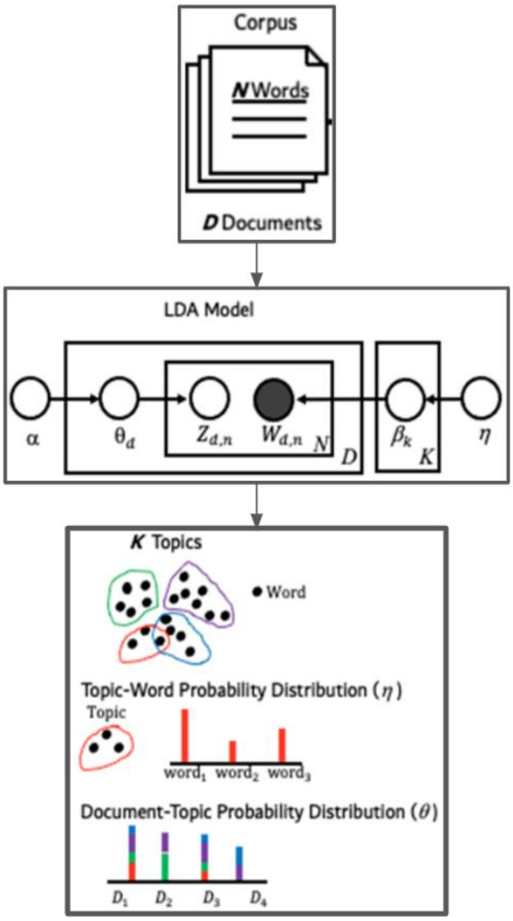

Fig 1. Graphical model of LDA topic modelling (adopted by (Blei, 2012; Hwang & Cho, 2021)).

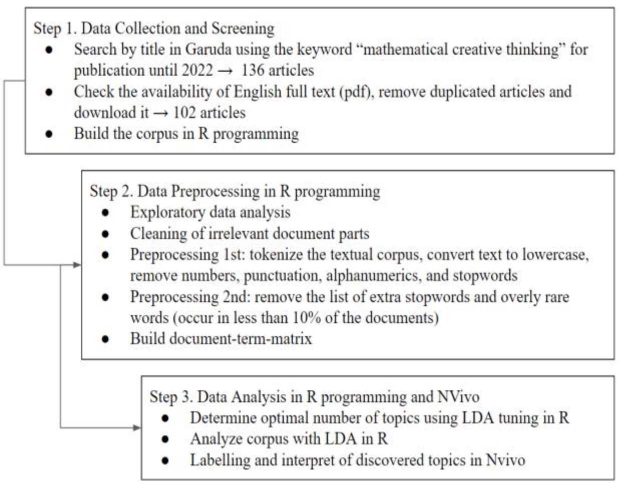

Fig 2. An overview of research metholodology.

This study uses LDA topic modeling, an unsupervised learning technique without data training to determine the best model. One of the essential parameters of LDA topic modeling is the number of topics. To determine the optimal topic number, (Vayansky & Kumar, 2020) and (Yang et al., 2018) recommended iteratively modeling the topic for the same data using different values of k and then analyzing specific metrics to compare the results. The establisment of the fitting method and alpha parameter also determines the modeling results. Several techniques, including the variational approach, expectation propagation, and Gibbs sampling, have been proposed to estimate LDA parameters (Jelodar et al., 2019).

The research stages were designed adopted by (Hwang & Cho, 2021) and (Gürcan, 2018) methods, consisting of data collection and screening, preprocessing, and analysis. The R programming language and RStudio for Windows were used to process and analyze the data. The stages are showed in Fig. 2.

2.1 Data collection and screening

The selected data source is the Garuda database (Garda Rujukan Digital), a platform for information sources for scientific publications in Indonesia covering all fields of knowledge. The search keyword used is "mathematical creative thinking" in the title (search by title), with the year of publication being limited to 2022. A total of 136 articles from 2010 - 2022 were obtained from the search. The articles collected are only in English. Next, check the availability of complete documents in pdf format and download them, if any. Duplicate articles are removed from the dataset. Due to their failure to meet the selection criteria, 34 articles were eliminated.

Finally, 102 articles were gathered as a corpus. For exploratory data analysis, metadata of each article, including the publication year, author name, article title, citation, and publication source, is gathered and kept in a Google worksheet. Each article is saved and given a name based on the year and publication source to ensure clarity.

2.2 Data preprocessing

The unique characteristic of text data is structural irregularity and a high proportion of noise. The (Debortoli et al., 2016) suggest two processes for data preparation. First, exploratory data analysis (EDA) to get an initial description of the dataset. EDA is the crucial process of doing preliminary data analyses to find patterns, identify anomalies, test hypotheses, and double-check assumptions with the aid of summary statistics and graphical representations.

The second process of preparing data is perform text cleaning and preprocessing. Quantitative analysis requires converting text into numeric data, where preprocessing steps are needed to build the relevant document-term-matrix (DTM) (Denny & Spirling, 2018). This study only focuses on English text in the title, abstract, and introduction until the conclusion. Therefore, before carrying out text preprocessing, Indonesian language abstracts, acknowledgments, references, and certain parts of the article considered irrelevant are deleted. Removal is done with the help of the Nitro Pro application. After cleaning is complete, proceed with text preprocessing.

Text preprocessing steps include tokenization, lowercase conversion, removing numbers and punctuation, removing stop-words, removing missing and misspelled words, and stemming (Debortoli et al., 2016; Denny & Spirling, 2018). The stemming process is essential to improve the ability to interpret LDA results. However, in several studies, stemming may not be done to avoid loss of rationalization due to the many special terms for a domain, as was done by (Gürcan, 2018).

This research uses the LDA function in R, preparing an R library that supports LDA topic modeling and visualization is also necessary. In addition, beside installing the topicmodels package, it also requires the pdftools library to read and process PDF files, tm to perform text mining, ldatuning to determine the optimal number of topics, tidytext to process data neatly, ggplot2 to display data, and dplyr to manipulate data. In tidytext, each variable is a column, each observation is a row, and each type of observation unit is a table, according to the concept of data cleansing. In a further cleaning stage, pdf files are read on R-studio, and if any are discovered unreadable by the library, they are removed from the data folder. All pdf documents are gathered into a corpus so that the topic modeling method may be carried out in R. Based on this corpus, DTM is subsequently created.

2.3 Data analysis

Topic models are generative models used in machine learning and natural language processing that offer a probabilistic framework for the term frequency occurrences in documents in a given corpus (Hornik & Grün, 2011). LDA topic modelling is unsupervised; thus, the number of topics is predetermined or chosen using specific metrics (Yang et al., 2018). This research used LDA tuning technique based on four metrics from the R package, as worked by (Yang et al., 2018). Arun2010 and CaoJuan2009 metrics for finding minimum extremes. Deveaud2014 and Griffiths2004 metrics for determining maximum extremes. Metrics Deveaud2014 and Griffiths2004 generally grow with the number of themes, whereas Arun2010 and CaoJuan2009 generally drop with the number of topics (Yang et al., 2018). The results of LDA tuning were analyzed to determine the optimal number of topic. Then proceed with processing dtm to produce topic models.

In this research, LDA is implemented using variational expectation-maximization (VEM) algorithm as a fitting method. VEM algorithm is for the LDA full Bayesian model which proposed by David Blei. In order to maximize the fit to the observed data, VEM adopts a parametric approximation to the posterior distribution of both parameters and additional latent variables (Jelodar et al., 2019). As illustrated in Fig. 1, the hyperparamater alpha, α, determines the topic combination (Blei, 2012). A higher α value causes the topic distribution to be pushed away from the simplex extremities resulting in "smoothing" and $\alpha < 1$ is suggested to create a bias towards sparsity (Steyvers & Griffiths, 2007; Vayansky & Kumar, 2020). Therefore, this experiment assigns a value of $\alpha = 0.024$ for LDA with VEM.

This study uses the R package topicmodels and the LDA() function (Hornik & Grün, 2011). The topics obtained are visualized, labeled, and interpreted according to the research objectives. For this reason, two stages of analysis were carried out. First, analyze the distribution of topics in the document using theta distribution. The variable θ is the set of latent variables to be concluded (Blei, 2012; Steyvers & Griffiths, 2007). Second, analyze the distribution of words in the topic using the beta distribution. The hyperparameter β can be interpreted as a count of previous observations on the number of times a word is sampled from a topic before any word from the corpus is observed (Blei, 2012; Steyvers & Griffiths, 2007).

The results of the beta analysis are mapped to the sample documents obtained from the theta analysis. To label the topics, a qualitative content analysis was carried out on the sample documents by codifying the sentences in the sample documents containing the keywords of a topic. The final analysis is carried out qualitatively in Nvivo, referring to the theoretical basis relevant to MCT.

3. Results and Discussion

The text corpus in this study came from 102 documents obtained from a search on one database, Garuda, for 2010 - 2022. However, the publication sources come from 55 English-language Indonesian journals and proceedings, as discussed in the exploratory data analysis section below. In this study, data are the full-text articles have been cleaned from irrelevant sections such as acknowledgment and references. In addition, the total data in the corpus is more than 150 thousand words after preprocessing. It contrasts with (Hwang & Cho, 2021), which employed the abstracts of 3485 articles and a corpus of about 1000 frequency of words. It means that a large corpus is available in this study to create topic models automatically.

3.1 Exploratory data analysis (EDA)

This study does EDA with the aid of bibliometric data to gain a general understanding of the dataset. The EDA comprises the distribution of publication sources (journals or proceedings) and the number of articles per year. The graph in Fig. 3 below demonstrates an upward trend in Indonesian MCT publications.

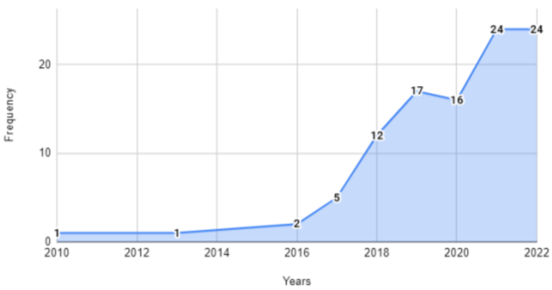

Fig 3. Growth of MCT publications in Indonesia per year

Since 2018, the topic of MCT has attracted significant interest from researchers in Indonesia. Even though it had dropped in 2020, during the 2021 - 2022, quantitatively, MCT publications in Indonesia increased rapidly. Fig. 3 also shows that in 2011, 2012, 2014, and 2015, no MCT publications in Indonesia were recorded in the Garuda database. During 2010 - 2016, MCT studies have not spread broadly in Indonesia. Two MCT articles have over 100 citations as of 2022, both were published in 2010 and 2016 respectively.

First, the article entitled "Leveling students' creative thinking in solving and posing mathematical problems" (Siswono, 2010) has been cited 177 times. Based on features of fluency, flexibility, and originality in posing and solving mathematical problems, the research produced a framework for the level of student's creative thinking in mathematics. This framework has been referred to by Indonesia researchers in (Nufus & Duskri, 2018) and abroad in (de Vink et al., 2022).

Second, an article entitled "Open-Ended Approach: An Effort in Cultivating Students' Mathematical Creative Thinking Ability and Self-Esteem in Mathematics" by (Fatah et al., 2016) has been cited 138 times. Their research results showed that the ability of MCT increased after giving treatment in the form of an open-ended approach. The research of the (Fatah et al., 2016) is adopted by (Hadar & Tirosh, 2019) to develop an analytic framework for critical thinking in the mathematics curriculum. (Siswono, 2010) and (Fatah et al., 2016) is considered a pioneering article in the field of MCT in Indonesia and has received international recognition.

As a result of the investigation into publication sources, there are 49 indexed national journals and six international conference proceedings. Table 1 presents 12 sources that have more than one article.

Table 1. Top publication sources of MCT in Indonesia

SOURCE'S TITLE AND LINK	TOTAL ART.	SINTA
Unnes Journal of Mathematics Education https://journal.unnes.ac.id/sju/index.php/ujme	16	Sinta 3
Unnes Journal of Mathematics Education Research https://journal.unnes.ac.id/sju/index.php/ujmer/index	10	Sinta 4
Journal of Innovative Mathematics Learning https://journal.ikipsiliwangi.ac.id/index.php/jiml/index	8	Sinta 4
Al-Jabar: Jurnal Pendidikan Matematika http://ejournal.radenintan.ac.id/index.php/al-jabar/index	6	Sinta 2
Journal on Mathematics Education https://ejournal.unsri.ac.id/index.php/jme	3	Sinta 1
Infinity: Journal of Mathematics Education http://e-journal.stkipsiliwangi.ac.id/index.php/infinity	3	Sinta 2
Kreano: Jurnal Matematika Kreatif-Inovatif https://journal.unnes.ac.id/nju/index.php/kreano/index	3	Sinta 2
Jurnal Pendidikan Matematika Indonesia https://journal.stkipsingkawang.ac.id/index.php/JPMI	2	Sinta 4
Mathematics Education Journal https://ejournal.umm.ac.id/index.php/MEJ/index	2	Sinta 4
Jurnal Cendekia: Jurnal Pendidikan Matematika https://j-cup.org/index.php/cendekia	2	Sinta 3
Journal of Primary Education https://journal.unnes.ac.id/sju/index.php/jpe	2	Sinta 3
Journal of Research and Advances in Mathematics Education https://journals.ums.ac.id/index.php/jramathedu/index	2	Sinta 2

Table 1 shows the top four journals that discuss MCT a lot, published by Semarang State University (Unnes). The journals are Unnes Journal of Mathematics Education, Unnes Journal of Mathematics Education Research, Kreano: Jurnal Matematika Kreatif-Inovatif, and Journal of Primary Education. According to the EDA above, the corpus may be considered suitably representative of the growth of MCT research in Indonesia between 2010 and 2022.

3.2 Text preprocessing

Text preprocessing is necessary to remove undesirable elements from the dataset and prepare it for subsequent processing. Table 2 compares the 20 words with the highest frequency before and after preprocessing.

Table 2. The 20 most common terms' frequency before and after preprocessing

STAGES	WORDS AND FREQUENCIES
Initial	the (22975), and (8690), that (4755), students (4724), learning (4321), thinking (4271), creative (3996), with (3277), mathematical (2870), are (2338), ability (2267), this (2233), can (1959), was (1795), for (1502), research (1433), results (1422), based (1386), from (1372), not (1309)
Preprocess 1st	student (8055), learn (5286), think (4899), creative (4316), ability (3475), problem (3229), mathematical (3006), test (2139), result (2098), solve (1716), research (1647), mathematics (1537), base (1447), skill (1414), class (1224), study (1209), subject (1172), model (1160), datum (1083), level (1072)
Preprocess 2nd	student (8055), learn (5286), think (4899), creative (4316), ability (3475), problem (3229), mathematical (3006), test (2139), result (2098), solve (1716), mathematics (1537), base (1447), skill (1414), class (1224), study (1209), subject (1172), model (1160), level (1072), answer (1065), question (1003)

Preprocessing was done twice for this investigation. The first preprocessing entails changing the text's case to lower case, getting rid of all numbers, getting rid of punctuation characters, lemmatizing with the lexicon, and getting rid of words that appear too frequently but are unnecessary based on the list of stopwords for English provided by snowball and smart in R. The frequency of words in the corpus are observed to identify any additional stopwords not listed in R's stopwords package. The second preprocessing stage was to delete words based on the list of additional stopwords and all words that appeared less than 10% in the corpus.

Table 2 illustrates how removing stopwords and stemming affects the occurrence and frequency of a word in the corpus. For example, the words "the", "and", and "that" which initially existed after the first preprocessing, disappeared from the corpus because they were included in the list of English stopwords provided in the R library. Meanwhile, the words "research" and "datum" still appearing in the first preprocessing stage were removed from the corpus in the second preprocessing by using an additional list of stopwords created by the researcher. The two words are considered irrelevant to the purpose of topic modeling, so they need to be removed from the corpus. An example of the results of lemmatization is that the words "learning" and "thinking" are changed to "learn" and "think" in the first preprocessing. However, the words "mathematics" and "mathematical" are retained because these two words have different contexts. The results of preprocessing can increase the effectiveness and efficiency of text processing in the next stage because it reduces the number of word variations and the total frequency of words that must be processed. The impact of preprocessing can also be seen from the number of words for each processed document. The graph below shows the change in the total frequency of words per document before and after preprocessing.

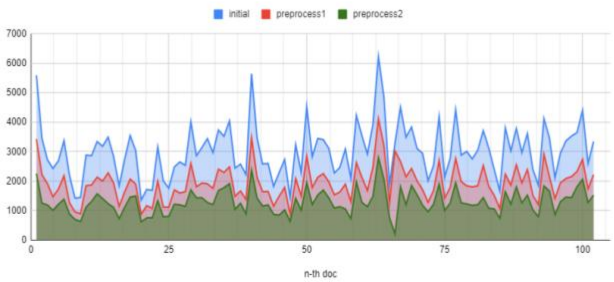

Fig 4. Changes in the total frequency of words per document before and after preprocessing.

Table 3. Comparison of statistics before and after preprocessing

STATISTICS	INITIAL	PREPROCESSING	
		First	Second
Total word frequency	317,425	202,768	156,288
Number of unique words	43,453	19,340	822
Sparsity	97%	97%	67%
Non-entries	112,853	62,030	27,424
Sparse entries	4,319,353	1,910,650	56,420

As shown in Fig. 4 illustrated that preprocessing does not change the word composition of each document. It only reduces irrelevant words to the topic modeling objective.

The preprocessing method considerably alters the total word frequency, the number of unique words, sparsity, empty entries, and sparse entries. DTM organizes all word occurrences in the corpus according to the document. The DTM divides terms (or words) into columns and documents into rows. The weighting process is done based on word frequency. Suppose a word appears in a specific document. The matrix entry for the row and column in that situation is 1. Otherwise, it is 0 (many occurrences within a document are recorded; for example, if a word appears twice, "2" is recorded in the appropriate matrix entry). In this study, the dtm created after the second preprocessing consisted of 102 lines (number of pdf documents) and 822 columns (number of unique words). Table 3 summarizes the statistics before and after preprocessing.

One of the changes in the total column is caused by the sparsity parameter. After deleting all words that appear less than 10%, the sparsity value decreases from 97% to 67%. The impact of the change in sparsity can also be seen from the significant decrease in total non-/sparse entries. The results shown in Table 3 are relevant to (Denny & Spirling, 2017) study, which demonstrated preprocessing impacts every input vocabulary and document used in the modeling procedure.

3.3 LDA tuning results

LDA tuning determine the optimal number of topics as input for LDA modeling. The appropriate metrics values are presented in Fig. 5 and Table 4 by changing the topic numbers from 15 to 25. A graphical visualization of the LDA tuning is shown in Fig. 5.

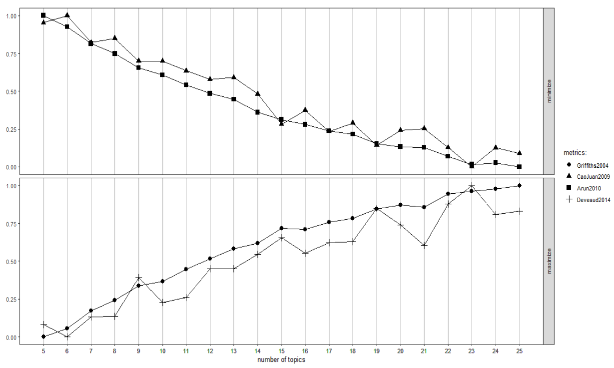

Fig 5. Visualization of LDA tuning results

The results of LDA Tuning are the optimal number of topics, k, with related metrics, as shown in Table 4.

Table 4. Results of LDA tuning metrics

k	Griffiths2004	CaoJuan2009	Arun2010	Deveaud2014
25	-564360	0.271263	114.0049	2.092403
24	-566066	0.281036	116.5704	2.081621
23	-567249	0.247492	115.5218	2.170705
22	-568534	0.281284	120.6377	2.113735
21	-575228	0.314578	126.165	1.986975
20	-574243	0.311516	126.6377	2.050591

The values in tables and graphs show that the Griffiths2004 and Arun2010 metrics suggest the optimal number of topics is 25, while the CaoJuan2009 and

Deveaud2014 metrics recommend 23 topics, respectively. Therefore, this study determines the number of topics to be used in the next stage is 23 topics. Furthermore, LDA topic modeling was carried out using the VEM method and alpha = 0.024 for k = 23 topics.

3.4 Distribution topics in documents

The theta (θ) distribution generates by LDA as the posterior probability of a topic in each document which states the topic proportions for a document (Blei, 2012). This study analyzes the distribution of topics in individual documents by presenting the number of documents mapped to 23 topics for varying theta values.

Table 5. The theta value for k = 23 topics

MINIMAL THETA (θ)	TOTAL DOC	% TOTAL DOC	MAX NUMBER OF TOPICS PER DOC
0.20	102	100.00%	3
0.25	101	99.02%	2
0.50	71	69.61%	1
0.75	42	41.18%	1

The experimental results show that setting the minimum value of theta affects the number of documents needed to describe the topic. The greater the theta value, the more specific the document is relevant to a topic. As shown in the Table 5, if the minimum theta is 0.20, there are at most three topics in a document. Conversely, if the theta value is more than or equal to 0.75, then only 42 documents are needed where at most, one topic is relevant to a document.

Although only about 41% of the total data, the sample selected based on the theta ≥ 0.75 covers all topics. The document is said to be relevant to the topic if the theta probability is greater than or equal to 0.75. Finding the documents that are most pertinent to a topic requires selecting document samples. The following graph shows the total of sample documents for each topic.

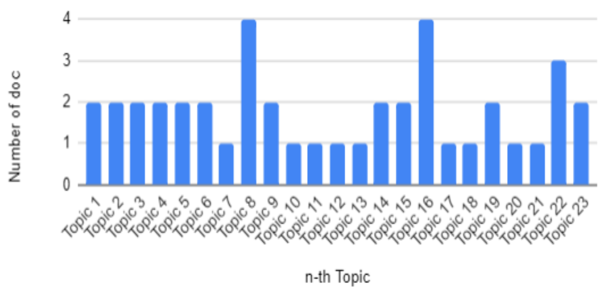

Fig 6. Total of sample documents for each topic.

The graph in Fig. 6 shows nine relevant topics in only one document, eleven topics are spread across two documents, only one topic is contained in three documents, and two topics are in four documents. As an illustration, based on the theta value, the relevant documents are selected for seven topics, as seen in below.

Table 6. Sample documents of the topics based on θ values

k	TITLE'S DOCUMENTS	θ
1	Improving the mathematical creative thinking ability of elementary students through the CRH learning model assisted by monopoly game media	0.89
	Application of Socrative-Based Assessment to Measure Mathematical Creative Thinking Ability with Problem Based Learning Model Viewed from Curiosity	0.95
2	Prospective teachers conception of mathematical creative thinking	0.76
	The creative thinking process of teachers in designing mathematical tasks	0.93
3	Guided discovery worksheet for increasing mathematical creative thinking and self-efficacy	0.94
4	The creative thinking ability in anchored instructions (AI) learning reviewed from mathematical disposition	0.96
	Application of Socrative-Based Assessment to Measure Mathematical Creative Thinking Ability with Problem Based Learning Model Viewed from Curiosity	0.99
5	Developing web-assisted interactive media to improve mathematical creative-thinking ability	0.97
	Developing and playing geometric puzzle game to enhance the ability of mathematical creative thinking	0.87
6	Mathematical issues in two-dimensional arithmetic for analyze students' metacognition and creative thinking skills	0.99
	Students' mathematical lateral thinking skills in creative problem-solving	0.96
7	Development of higher order thinking skills (HOTS) assessment instruments to improve students' mathematical creative thinking skills	0.99
8	Improving student's mathematical creative thinking ability and self regulated learningusing sylver approach	0.84
	Mathematical creative thinking and habits of mind grounded on student's cognitive stage	0.98
	The effect of exploration approach on students' mathematical creative thinking ability and disposition	0.99
	The effect of inductive-deductive approach on students' mathematical creative thinking ability and self efficacy	0.99
9	Analysis of Students' Mathematical Creative Thinking Ability in Module-assisted Online Learning in terms of Self-efficacy	0.99
	Flipped Classroom Learning: Mathematical Creative Thinking Skills Based on Mathematical Resilience using Augmented Reality	0.86
10	Description of Mathematical Creative Thinking Ability of High School Students viewed from Metacognition Skills	0.89
11	Systematic Literature Review: STEM Approach through Engineering Design Process with Project Based Learning Model to Improve Mathematical Creative Thinking Skills	0.84
12	Analysis of Mathematical Reasoning Ability and Mathematical Creative Thinking Elementary School Students in Solving Story Problems	0.95
13	The influence of PJBL-Stem and PBL-based on the learning motivation of the students in the mathematical creative thinking skills	0.99
14	Leveling students' creative thinking in solving and posing mathematical problem	0.99
	Mathematical Creative Thinking Leveling on NonMathematics Department Students	0.78
15	Description of Mathematical Creative Thinking and Reasoning Ability of SMP Students in Islamic Culture-Based Learning	0.93
	Developing character based interactive learning media to facilitate students' self-learning of mathematics capita selecta	0.98

k	TITLE'S DOCUMENTS	θ
16	The students' mathematical creative thinking ability of junior high school through problem-solving approach	0.86
	Discovery Learning with Scaffolding To Promote Mathematical Creative Thinking Ability And Self-Efficacy	0.85
	Advocacy approach with open-ended problems to mathematical creative thinking ability	0.78
	Discovery Learning Versus Traditional Learning: How Effective Discovery Learning Can Improve Mathematical Creative Thinking Skills	0.86
17	Literature Review: Mathematical Creative Thinking Ability, and Students' Self Regulated Learning to Use an Open Ended Approach	0.85
18	The mathematical creative thinking ability viewed from learning interest in eleventh grade of vocational high school by using treffinger model assisted by problem card	0.83
19	Analysis of Student's Creative Thinking Ability in Mathematical Problem Solving in Terms of Extrovert and Introvert Personality Types	0.82
	Profile of Students' Mathematical Creative Thinking Ability in Solving Mathematical Problem	0.98
20	Mathematical Creative Thinking Ability of Junior High School Students in Solving Open-Ended Problem	0.78
21	Convergent and divergent thinking in mathematical creative thinking processes in terms of students' brain dominance	0.99
22	The effect of students' leaning style and self-concept toward mathematical creative thinking ability in linear algebra subject	0.85
	Discovery and core learning model toward creative thinking viewed from logical mathematical intelligence	0.99
	CORE Teaching Model Based Mnemonic Technique Impact Students' Mathematical Creative Thinking Ability and Metacognitive Awareness	0.98
23	Developing integration techniques module to improve mathematical creative thinking ability in integral calculus	0.88
	The development of mathematics learning devices based on realistic approach to improve creative thinking and mathematical communication ability for junior high school students at SMP PAB 2 Helvetia	0.99

At a glance, by simply observing the document titles, as shown in Table 6, it can be analyzed that there are similarities between some topics. For examples, topic 3, 5, and 23 relevant to learning resources of MCT. An analysis of the probability of words in a topic is carried out to dig deeper into the relationship between the topics. Therefore, the beta posterior probability is used to understand the distribution of words in a topic.

3.5 Distribution word in topics

This study analyzes scientific articles on MCT in Indonesia from indexed journals and proceedings at Garuda. The results of LDA modeling by specifying 23 topics provide a collection of topics represented by some keywords based on probability models.

LDA has a hyperparameter β which represents topic-word density and states the probability of a word being associated with a topic (Blei, 2012; Hwang & Cho, 2021). Each word in the corpus has the probability of contributing to a topic. This study limits the analysis based on the five words with the largest beta values and compares topics based on the total beta values of the five, as seen in Table 7.

Table 7. MCT research topics in Indonesia based on LDA topic modelling

k	TOP 5 KEYWORDS & β PROBABILITY	TOTAL β
1	learn (0.058), **student (0.057),** think (0.039), test (0.037), creative (0.033)	0.225
2	problem (0.057), think (0.053), idea (0.048), creative (0.044), **student (0.042)**	0.245
3	worksheet (0.052), **student (0.047),** test (0.034), creative (0.032), think (0.032)	0.197
4	**student (0.056)**, ability (0.043), think (0.042), creative (0.041), learn (0.038)	0.220
5	**student (0.066),** learn (0.039), develop (0.028), medium (0.027), base (0.019)	0.179
6	think (0.054), **student (0.054),** skill (0.052), level (0.035), creative (0.032)	0.227
7	assessment (0.059), ability (0.036), instrument (0.036), think (0.035), **student (0.033)**	0.199
8	**student (0.081),** think (0.028), grade (0.026), mathematical (0.024), approach (0.023)	0.182
9	**student (0.068),** learn (0.049), skill (0.022), creative (0.021), think (0.021)	0.181
10	**student (0.062),** style (0.038), problem (0.034), ability (0.032), answer (0.030)	0.196
11	learn (0.043), creative (0.038), think (0.036), **student (0.035),** skill (0.029)	0.179
12	**student (0.100),** problem (0.044), mathematical (0.039), ability (0.038), reason (0.025)	0.246
13	learn (0.097), **student (0.071),** motivation (0.042), think (0.030), skill (0.029)	0.268
14	problem (0.066), **student (0.058),** level (0.056), think (0.047), creative (0.033)	0.260
15	**student (0.053),** learn (0.046), think (0.026), mathematics (0.023), ability (0.019)	0.166
16	learn (0.069), **student (0.059**), class (0.034), ability (0.032), mathematical (0.030)	0.224
17	**student (0.072),** learn (0.057), think (0.043), creative (0.042), openended (0.040)	0.256
18	learn (0.087), **student (0.073),** think (0.051), ability (0.045), creative (0.044)	0.300
19	subject (0.085), creative (0.069), problem (0.052), solve (0.044), think (0.041)	0.291
20	**student (0.090),** problem (0.062), ability (0.049), think (0.042), creative (0.035)	0.278

k	TOP 5 KEYWORDS & β PROBABILITY	TOTAL β
21	think (0.083), **student (0.065),** question (0.051), process (0.043), creative (0.040)	0.281
22	learn (0.063), **student (0.048),** model (0.046), think (0.039), creative (0.036)	0.232
23	learn (0.041), material (0.041), teach (0.032), **student (0.030),** test (0.027)	0.171

An analysis of the occurrence and probability of words on a topic presented in the table 7 concluded that 22 topics contained the top keyword "student", except for the 19th topic labeled "creative problem solving". Evaluating all the words in the corpus (822 unique words), the word "student" has a frequency of 8,055, which is about 5.15% of the total frequency of 156,288 (see Tables 2 and 3). The range of "student" beta probabilities on all topics is quite broad, between 0.03 to 0.1. The topic has more words when the beta value is higher; conversely, when the beta value is lower, the topic contains fewer words and more specific words (Blei, 2012). This means that the word "student" has become very general, so contextually, it can be understood that MCT research in Indonesia focuses on students. Therefore for labeling purposes, the word "student" can be omitted.

Table 7 also presents that two or more topics show the similarity of keywords. This means the topics refer to the same concept. Therefore, a qualitative content analysis was carried out on the documents contributing to these topics to obtain the MCT dimensions and descriptions. Overall, there are 33 keywords spread across 23 topics. Besides "student", there are nine other keywords that appear on more than one topic.

Researchers visualize topics with LDAvis developed by (Sievert, & Shirley, 2014) to facilitate deeper analysis. The topic model is shown in the left panel, with topics represented by circles in a two-dimensional plane. The center is located by measuring the separation between topics and mapping that separation onto two dimensions using a multidimensional scale. The following is an example of visualization for topic 1.

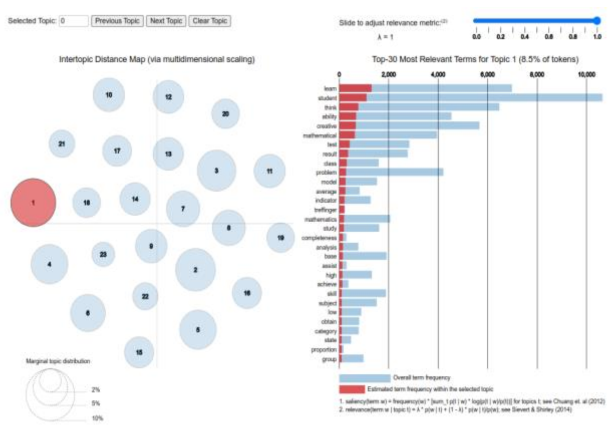

Fig 7. Example of LDAvis for topic of MCT

In this visualization, the frequency of words on a topic could be compared to the overall frequency in the corpus. On the other hand, LDAvis can also be used to analyze the presence of a word in various topics. The right panel shows a horizontal bar chart with the bars representing the specific terms most helpful in understanding the chosen topic.

The results of the analysis of the visualizations on LDAvis were used to compare the frequency percentage and the number of topics for specific keywords, as described below. Fig. 8 shows the words "think" and "creative," both in terms of frequency and distribution of topics, which are common in MCT research. This is understandable, the two phrases as search keywords for articles in the Garuda database. For topic labeling, both words may be omitted.

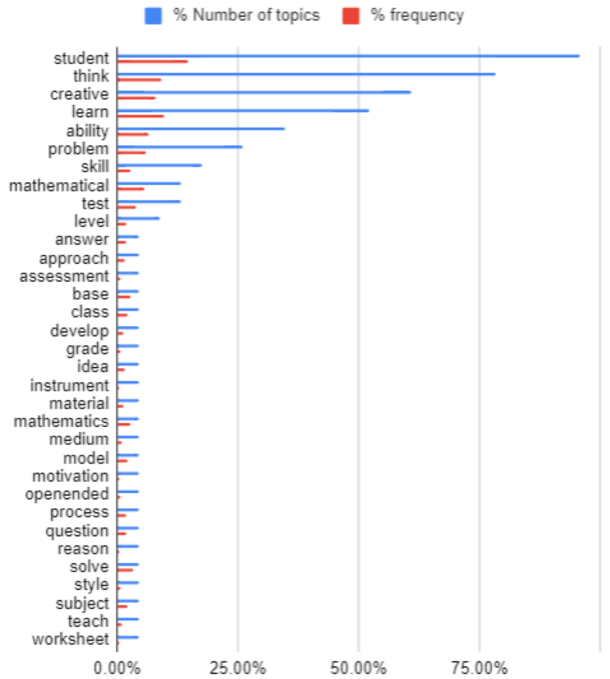

Fig 8. The frequency percentage and the number of topics for specific keywords.

The next example illustrates how to use Table 6, Fig. 6 and Fig. 7 to conduct deeply topic analysis. According to beta probability value, 18th topic, which includes the words "learn," "student," "think," "ability," and "creative," has the biggest overall beta of 0.30. The lowest overall beta is 0.166 for subject 15, which contains the words "student," "learn," "think," "math," and "ability." Topic 18th has more common terms, as seen by a larger beta value. In contrast, topic 15th contains fewer general words.

The distribution of topics within the documents is then analyzed to obtain a description of each category label. NVivo was utilized to conduct a content analysis on the example documents. The five categories of topic are described as a result of the qualitative content analysis. The category description refers to the important phrases for each topic that may be found in the sentences of the sample document. Hence, labeling concentrates on particular terms and reference instances of sentences in documents that contribute to each topic.

Tabel 8. Label of topics

LABELS	TOPICS	EXAMPLE OF SENTENCES IN SAMPLE DOCUMENTS
Learning resources (5 topics)	3, 5, 15, 23	Having encountered this **problem,** the **media** game was one alternative way of **learning** to improve **students' mathematical creative thinking skills** (Susiaty, 2021). Many **students** looked bored and did not understand what the teacher explained, so this affected **student learning** outcomes which resulted in low **creative thinking ability** and **student's mathematical** communication (Wiraprana & Surya, 2022).
Learning approach (6 topics)	8, 11, 13, 16, 17, 22	One of the appropriate **learning approaches** to addressing the **problem** of **mathematical creative thinking skills** and self-regulated **learning** is the **Open-Ended approach** (Ali et al., 2021). Model Eliciting Activities (IDA) gave bigger role than SA on improving **students' mathematical creative thinking ability** (MCTA) it gains and on MSE. **Students** getting treatment with IDA attained MCTA and MSE at successively pretty good and good **grades level**. While students taught by scientific **approach** (SA) obtained at low-medium **grade level** (Mahmudin, 2020).
Assessment (2 topics)	1, 4, 7	To see **students' mathematical creative thinking abilities based** on pretest and posttest results, researchers used an **assessment** rubric based on aspects of **students' mathematical creative thinking** developed by Bosch (1997) with indicators of fluency, flexibility, originality, and elaboration (Arifuddin et al., 2022). One of the causes of PBL model by applying Socrative-based **assessment** to be effective towards the achievement of **students' mathematical creative thinking ability** is that PBL model begins by giving real **problems** to **students**, thus helping **students** in the **learning** process before knowing formal concepts (Sari & Masrukan, 2022).
Teacher's Competencies (1 topic)	2	The prospective teacher's conception of the **creative thinking** of **mathematics** leads to the emergence of new concepts or **ideas** in completing **mathematical problems** based on experience (Purwosetiyono et al., 2018). In order to encourage **students** in **thinking** and reasoning **mathematical ideas**, it is important for teachers to focus on **creative** tasks (Taufiq et al., 2018)
Student's competencies (9 topic)	6, 9, 10, 12, 14, 18, 19, 20, 21	If someone has the **ability** to **think creatively**, they will find it easier to **solve problems** in various ways both **problems** in real life or in **mathematics** (Daiana et al., 2021). Several factors affect the **level** of **students**' metacognition and **creative thinking skills**, namely **understanding** of information, **skills** in choosing strategies, developing the right strategy, elaborating **answers**, mastery of **mathematical material**, and the tendency to rely on memorization (Tohir & Muhasshanah, 2021).

Thus, the 23 topics obtained from LDA modeling can be grouped into five labels: learning approach, assessment, learning resources, teacher's competencies, and student's competencies. Each category label comprises some specific topics, as indicated in Table 8. For instance, three topics are included in the assessment category of MCT research with specific keywords: learn, ability, learn, assessment, ability, instrument. Nine topics made up of words: skill, level, learn, style, problem, ability, answer, problem, mathematical, reason, subject, solve, and question are included in the student's competency. The learning approach for MCT is the following category, which includes six subjects under the headings grade, mathematical, approach, learn, skill, motivation, class, ability, openended, model. A worksheet, test, learn, develop, medium, base, mathematics, ability, material, teach are keywords of topics 3, 5,15, and 23 that are included in MCT's fourth category of learning resources. The fifth label is teacher competencies which only appears in topic 2 with the specific words problem and idea. Sample documents relevant to this topic are entitled "Prospective teachers' conception of mathematical creative thinking" (Purwosetiyono et al., 2018) and "The creative thinking process of teachers in designing mathematical tasks" (Taufiq et al., 2018). The labeling is not only based on the word but the context of the word in the sentence in each example document, as presented in Table 8. The topic labelling is comparable to (Hwang & Cho, 2021).

4. Conclusion

This research applies LDA topic modeling to identify latent topics in scientific articles about MCT in Indonesia. Two posterior probabilities of LDA, namely beta and theta, are used to explore a topic's essential words and the document's contribution to a topic. The results of this study indicate that there are 23 MCT research topics in Indonesia which can be grouped into five categories which are the focus of MCT research in Indonesia. The five categories are learning resources, learning approaches, assessment, student competencies, and teacher competencies. The impact of modeling this topic was identifying MCT research topics in Indonesia during the 2010 - 2022 range. In addition, a research gap was also found, namely that there was still little research that focused on improving teacher competence. These findings indicate that LDA topic modeling can also be used to formulate state-of-the-art in a research field.

Conflict of interest

The author affirms that have no known financial or interpersonal conflicts that would have appeared to have an impact on the research presented in this scientific paper.

References

Ali, D., Zubaidah Amir, M. Z., & Vebrianto, R. (2021). Literature Review: Mathematical Creative Thinking Ability, and Students' Self Regulated Learning to Use an Open Ended Approach. *Malikussaleh Journal of Mathematics Learning,* 4(1), 52-61. http://doi.org/10.29103/mjml.v4i1.3095

Arifuddin, A., Wahyudin, W., Prabawanto, S., Yasin, M.,

& Elizanti, D. (2022). The Effectiveness of Augmented Reality-Assisted Scientific Approach to Improve Mathematical Creative Thinking Ability of Elementary School Students. *Al Ibtida: Jurnal Pendidikan Guru MI*, 9(2), 444-455. http://dx.doi.org/10.24235/al.ibtida.snj.v9i2.11647

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84. http://doi.acm.org/10.1145/2133806.2133826

Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of Computers & Education. Computers & Education, 151, 103855. https://doi.org/10.1016/j.compedu.2020.103855

Daiana, P., Surahmat, S., & Fathani, A.H. (2021). Profile of Students' Mathematical Creative Thinking Ability in Solving Mathematical Problem. *Formatif: Jurnal Ilmiah Pendidikan MIPA*, 11(1), 49-62. http://dx.doi.org/10.30998/formatif.v11i1.7810

Debortoli, S., Müller, O., Junglas, I., & Vom Brocke, J. (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, 39(1), 7. https://doi.org/10.17705/1CAIS.03907

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2), 168-189. https://doi.org/10.1017/pan.2017.44

De Vink, I. C., Willemsen, R. H., Lazonder, A. W., & Kroesbergen, E. H. (2022). Creativity in mathematics performance: The role of divergent and convergent thinking. *British Journal of Educational Psychology*, 92(2), e12459. https://doi.org/10.1111/bjep.12459

Fatah, A., Suryadi, D., & Sabandar, J. (2016). Open-Ended Approach: An Effort in Cultivating Students' Mathematical Creative Thinking Ability and Self-Esteem in Mathematics. *Journal on Mathematics Education*, 7(1), 11-20. https://doi.org/10.22342/jme.7.1.2813.9-18

Gürcan, F. (2018, September). Major research topics in big data: A literature analysis from 2013 to 2017 using probabilistic topic models. In 2018 *International Conference on Artificial Intelligence and Data Processing* (IDAP) (pp. 1-4). IEEE. https://doi.org/10.1109/IDAP.2018.8620815.

Hadar, L. L., & Tirosh, M. (2019). Creative thinking in mathematics curriculum: An analytic framework. *Thinking Skills and Creativity*, 33, 100585. https://doi.org/10.1016/j.tsc.2019.100585

Hornik, K., & Grün, B. (2011). topicmodels: An R package for fitting topic models. *Journal of statistical software*, 40(13), 1-30. https://doi.org/10.18637/jss.v040.i13

Hwang, S., & Cho, E. (2021). Exploring Latent Topics and Research Trends in Mathematics Teachers' Knowledge Using Topic Modeling: A Systematic Review. *Mathematics*, 9(22), 2956. https://doi.org/10.3390/math9222956

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211. https://doi.org/10.1007/s11042-018-6894-4

Mahmudin, C., Sumarmo, U., & Kustiana, A. (2020). The Effect of Inductive-Deductive Approach on Students' Mathematical Creative Thinking Ability and Self-Efficacy. *Journal of Innovative Mathematics Learning* (JIML), 3(4), 215-226. http://dx.doi.org/10.22460/jiml.v3i4.p215-226

Newton, D., Wang, Y., & Newton, L. (2022). 'Allowing them to dream': fostering creativity in mathematics undergraduates. *Journal of Further and Higher Education*, 1-13. https://doi.org/10.1080/0309877X.2022.2075719

Nufus, H., & Duskri, M. (2018). Mathematical Creative Thinking and Student Self-Confidence in the Challenge-Based Learning Approach. *Journal of Research and Advances in Mathematics Education*, 3(2), 57-68. https://doi.org/10.23917/jramathedu.v3i2.6367

OECD. (2019). Draft framework for the assessment of creative thinking in Pisa 2021. Paris: OECD Publisher.

Purwosetiyono, F. X. D., Sa'dijah, C., Hidayanto, E., Chandra, T. D., As'ari, A. R., & Irawan, E. B. (2018). Prospective teachers' conception of mathematical creative thinking. *International Journal of Insights for Mathematics Teaching* (IJOIMT), 1(2), 106-114. ISSN 2087-8885.

Sari, D. N. K., & Masrukan, M. (2022). Application of Socrative-Based Assessment to Measure Mathematical Creative Thinking Ability with Problem Based Learning Model Viewed from Curiosity. *Unnes Journal of Mathematics Education*, 11(2),137-147. https://doi.org/10.15294/ujme.v11i2.59901

Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70). http://dx.doi.org/10.3115/v1/W14-3110

Siswono, T. Y. E. (2010). Leveling students' creative thinking in solving and posing mathematical problem. *Journal on Mathematics Education*, 1(1), 17-40. https://doi.org/10.22342/jme.1.1.794.17-40

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In Handbook of latent semantic analysis (pp. 439-460). Psychology Press.

Suherman, S., & Vidákovich, T. (2022). Assessment of Mathematical Creative Thinking: A Systematic Review. *Thinking Skills and Creativity*, 101019. https://doi.org/10.1016/j.tsc.2022.101019

Susiaty, U. D., Prihatin, I., & Hartono, H. (2021). Developing and Playing Geometric Puzzle Game to Enhance the Ability of Mathematical Creative Thinking. Kreano, Jurnal Matematika Kreatif-Inovatif, 12(1), 39-50. https://doi.org/10.15294/kreano.v12i1.26613

Syaputri, A. W., Irwandi, E., & Mustakim, M. (2020). Naïve Bayes Algorithm for Classification of Student Major's Specialization. *Journal of*

Intelligent Computing and Health Informatics (JICHI), 1(1), 15-19. https://doi.org/10.26714/jichi.v1i1.5570

Taufiq, T., Budiarto, M. T., & Siswono, T. Y. E. (2018, December). The creative thinking process of teachers in designing mathematical tasks. In *International Conference on Mathematics and Science Education of Universitas Pendidikan Indonesia* (Vol. 3, pp. 634-638). P-ISSN 2655-2361, E-ISSN 2655-3252.

Tohir, M., & Muhasshanah, M. (2021). Mathematical issues in two-dimensional arithmetic for analyze students' metacognition and creative thinking skills. *Alifmatika: Jurnal Pendidikan Dan Pembelajaran Matematika*, 3(2), 170-183. https://doi.org/10.35316/alifmatika.2021.v3i2.170-183

Wiraprana, S., & Surya, E. (2022). The Development of Mathematics Learning Devices Based on Realistic Approaches to Improve Creative Thinking and Mathematical Communication Skills for 8th Grade Junior High School Students. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, 6(2), 2011-2024. https://doi.org/10.31004/cendekia.v6i2.1443

Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T., & Coimbra, C. F. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, 168, 60-101. https://doi.org/10.1016/j.solener.2017.11.023

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. Information Systems, 94, 101582. https://doi.org/10.1016/j.is.2020.101582