Digital Zone
Jurnal Teknologi Informasi dan Komunikasi

Vol: 15 No 02 2024 E-ISSN: 2477-3255

Accepted by the Editor: 01-10-2024 | Revision: 08-11-2024 | Published: 30-11-2024

Evaluation of Creative Economy and Tourism Industry Trends based on LDA Analysis with BERTopic

Icha Nura Nugraha[1], **Ema Utami**[2]
[1,2]Magister of Informatics Engineering,
University of Amikom Yogyakarta, Yogyakarta, Indonesia.
[1,2]Jl. Ring Road Utara, Yogyakarta, Ngringin, Condongcatur, Kec. Depok, Kab. Sleman, 55283
e-mail: ichanugraha@students.amikom.ac.id, ema.u@amikom.ac.id

Abstrak

Creative economy and tourism industry have a role in contributing country's foreign exchange. Efforts continue to be improved by utilizing social media. Latent Dirichlet allocation (LDA) and BERTopic topic model are used as topic models for creative economy and tourism trend analysis. The evaluation was carried out using a coherence matrix, topic distribution, similarity, and topic identification over the last five-years period. BERTopic has a higher coherence value of 0.53 compared to LDA 0.30 although the number of outlier topics dominates. The identification of the most relevant main topic trends is finance, travel, beaches and investment. These themes are interrelated in driving the growth of the creative economy and tourism, which increases local income and innovation in related sectors. BERTopic identifies hidden topics such as bitcoin cryptocurrency. In contrast, LDA provides a more even distribution of topics, revealing traditional trends such as beach tourism and travel. The evaluation offers key recommendations on creative economy and tourism policies to innovations about investment.

Keywords: *Trends, Creative economy tourism, Topic model, LDA, BERTopic.*

1. Pendahuluan

Creative economy and tourism are sources of foreign exchange for the country and are also one of the sectors contributing to the country's Gross domestic product (GDP) [1][2]. The creative economy and tourism contribute to generating income by leveraging trade and intellectual property, creating various opportunities in the process [2]. The recovery of the national economy in the tourism and creative economy sectors following the Covid-19 pandemic presents a significant challenge for both private entrepreneurs and the government. In 2022, the tourism sector generated approximately USD 6.72 billion in foreign exchange, contributing 3.6% to the national GDP. This figure saw a significant increase in 2023, reaching an estimated USD 7.08 9.99 billion. In 2022, the creative economy achieved a value of 1,280 trillion IDR with exports worth USD 26.94 billion. This grew significantly in 2023 to 342.92 trillion IDR with exports reaching USD 26.46 billion, and efforts to further enhance these figures will continue [3].

https://doi.org/10.31849/digitalzone.v15i2. 21030

The growth of the creative economy and tourism can be observed through the increasing number of foreign tourist arrivals. This is evident from data recorded at various entry points, showing 798,469 visits in 2023 with a growth rate of 16.19%. This upward trend continued into January 2024, with 927,746 visits, reflecting an 18.07% increase based on mobile positioning records of border entry points [4] the hospitality, transportation, merchandise, food service and culinary industries will also enjoy the impact of this event.

Advancement in information and communication technology have transformed how we interact, access information and consume content. In 2023, population data shows that out of 8.01 billion people, 5.16 billion were internet users, with 4.76 billion actively using social media [5]. Unstructured data continues to grow every year. Twitter is a social media platform that produces unstructured data [6]. Research on trends towards a topic has been conducted to find out the trend patterns on the topic of new and renewable energy in Indonesia based on tweets on the social media Twitter, this topic was conducted with the help of the Latent Dirichlet Allocation method [6].

Research on trends to see certain patterns is always studied such as on topic trends that are carried out to identify business intelligence trend patterns for the last 20 years on a topic. This process identifies research on an article. Identification of the research topic, using Latent Dirichlet Allocation is able to support potential future directions [7]. Related research on recurring topic trends is conducted with better and more sophisticated computational development of language models. The study analyzes 31 years of trends in the field of language models. The study adopts BERTopic for analysis related to 13,751 language models [8].

BERTopic with Latent Dirichlet Allocation is used simultaneously to explore in news text analysis. The study presents a comparative analysis of the two models that can provide recommendations for the future [9]. Topic modeling is often used to identify and categorize a topic and idea. Topic modeling and clustering can help to correlate and benefit each other. Topic modeling extracts words from documents, but often the topics are not coherent [10]. Clustering applications are often used. The k-means algorithm is the most frequently applied clustering method [11] while the algorithm Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) development of DBSCAN. HDSCAN is the clustering used by default from BERTopic to group the result vectors of the BERT model.

Clustering powerful data mining, the right number of clusters is the optimal solution for many clustering problems [11] in outlier and noise datasets. Coherence enhancement can also take advantage of Bidirectional Encoder Representations derived from Transformers (BERT) which allows the architecture to extract trained representations of words and sentences so that it can apply feature tuning [12]. Research on topic modeling is able to overcome unstructured social media data sets, but the evaluation steps that occur do not inform whether the resulting topics produce meaningful insights for those who examine social media [13]

Efforts to overcome the problem in measuring the alignment between automatic and human evaluations need to be improved, the problem will recur if researchers do not build models optimally by exceeding the limitations of the models used. There are 189 articles found by researchers using sub-optimal topic models [13]. The use of questionnaires and visuals [14] in testing is used to conclude the model, able to provide results from a perspective that is not sourced from expert assessments. Assessment without measuring the topic model evaluation matrix becomes sub-optimal research results.

Building an LDA model is a common problem when deciding the hyperparameter values in building an optimal model. Researchers agree that there is no automatic setting that works well for any dataset [15]. Optimization and evaluation of the most common models as a benchmark for validating the performance of various coherence models $\boxed{C_{UMass}}$, $\boxed{C_{PMI}}$, $\boxed{C_{NPMI}}$, $\boxed{C_V}$ or a combination of perplexity and coherence scores is used [13]. Coherence is used to define how interpretable a topic is. An important aspect of the overall problem can be seen by optimizing topic coherence to find the optimal topic similarity threshold [16].

Building topic models as an analysis requires optimization in each proposed approach. Previous studies have shown that LDA tuned in the setting leads to sub-optimal performance [15]. Recommendation [13] research transparency in experimental settings, parameters up to preprocessing of topic models is required. Efforts to overcome these problems such as measuring the alignment between automated and human evaluations through a parameter tuning approach, using architecture and evaluating coherence matrices transparently reduce sub-optimal research. Optimization of parameter usage and architecture and clustering between LDA and BERTopic models is expected to improve the alignment of automatic and human evaluation based on the evaluation value of the coherence matrix.

Research studies on recurring trends sometimes focus only on one method without conducting exploration without using model parameter tuning which can produce low quality topic results [15] Studies that consistently show sub-optimal use do not fully measure the alignment of model results and understanding. Based on the background of the problems in the explanation above, we conducted comprehensive research on the Creative Economy and Tourism on social media based on the Latent Dirichlet Allocation model with BERTopic.

The research will provide optimal insight to evaluate the trends of the creative economy and tourism industry based on the analysis model. From this research, it will be known whether the model is able to provide trend patterns with the same topics or tends to be different from the creative economy and tourism industry trends to provide recommendations for future topics. We hypothesize trends regarding creative economy and tourism towards social media patterns. Trends that are always changing need to adapt to events at any time. The main objective of this research is the theoretical and applied contribution in experiments such as performance, preprocessing to limitation and evaluation of topic models.

1. Evaluation : what are the trending topics generated over the last 5 years.
2. Comparation : do topic model methods produce the same number of topic patterns.
3. Experiment setup: based on our experiments for subsequent researchers.
4. Trend Analysis: to conduct an analysis of identified research themes to understand the shift in research focus and emerging trends over the last 5 years.

This paper will be summarized as a sequence consisting of an introduction to the model theory, data, methods, experiments, evaluation and conclusions of our research.

2. Research methods

First, in general, the proposed research stages consist of several basic steps. These steps generally explain how the proposed method consists of Twitter data collection, preprocessing, feature extraction, topic modeling offered, and topic score analysis in a comprehensive experiment figure (1).

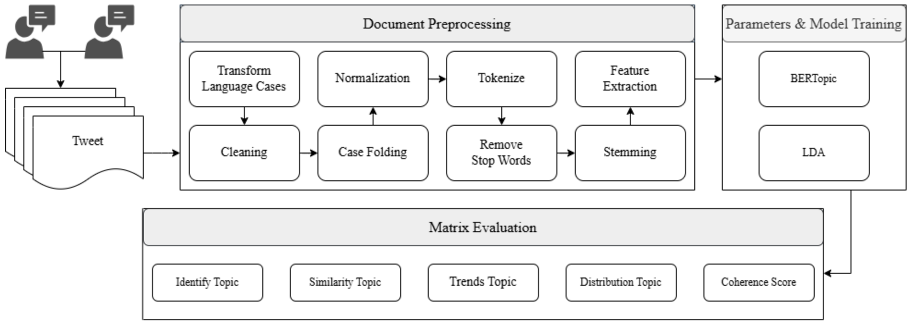


Figure 1. Sequence of steps for the proposed approach research

2.1. Method of Dataset Collecting

The API data provided by twitter is used to retrieve Tweet data. Data facilities based on certain keywords. The first collection process is to register a Twitter platform developer account. Twitter developers provide tokens that can be obtained and help the authentication process. The data retrieval process uses and accesses the Twitter API access library. The retrieval process involves several keywords with several time ranges taken every month in the last five years as follows, the keywords used, the details are shown in the table (1).

Table 1. Keywords Crawling and Result Detail Dataset

Keywords	Translated	Years	Data Tweet	Length Avg	Word Count Avg
Wisata	Tour	2019-2023	23165	147.79	19.1
Umkm	Umkm	2019-2023	18297	157.53	19.9
Turis	Tourist	2019-2023	24344	130.25	18.6

2.2. Preprocessing

Preprocessing is the process of preparing raw data into processed data. Recommendations [17] Preprocessing is done to follow the standards of the text obtained, arranged and analyzed statistically such as punctuation, capitalization, numbers, unnecessary spaces. Each step of this preprocessing aims to clean and organize the text into its original form, the purpose of preprocessing helps text analysis to be carried out and produce efficient and accurate text. Proper preprocessing is the key to obtaining meaningful and reliable analysis results from raw text data.

2.3. Topic Modeling Approaches

Topic modeling is a statistical method used to identify specific topics or patterns within a collection of documents [18][19] it is also a collection of text analysis techniques with the potential to enhance analytics, machine learning, and business intelligence in the near future [20] structural approach to managing large companies efficiently [21]. Latent Dirichlet Allocation is a technique used to uncover collections of texts in topic exploration, compile, organize, and ensure data is ready for analysis [14]. Latent Dirichlet Allocation is believed to be effective in generating probabilistic models on the topic [22].

Analyzing the relationship between a collection of documents that contain similar meanings that appear in the same text [22]. BERTopic [23][24] Inspired, several pre-trained Transformers have been proposed. [25] is a topic modeling technique used to train a dataset based on the latest pre-trained data that utilizes BERT dan c-TF-IDF in building topic interpretation clusters to retain important words from topic descriptions. Topic trends change over time, what can provide valuable content for researchers is the process of extracting information and meaning from semantic words [26][27]. The annual proportion of the number of topics reflects the level of attention on the topic. We conducted testing in this study. topic modeling, the object of comparison is LDA dan BERTopic.

Latent Dirichlet Allocation is one of the techniques that is widely used to reveal a collection of texts in the exploration of probabilistic generative model topics in a corpus [28]. Latent Dirichlet Allocation is believed to be effective in producing probabilistic models on topics. Latent Dirichlet Allocation can be represented by figures (2).

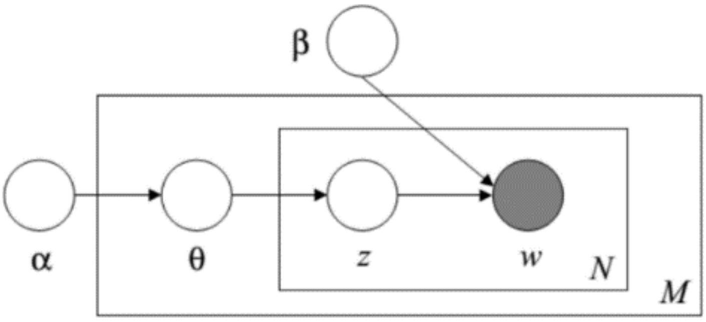


Figure 2. Graphical model Latent Dirichlet Allocation (LDA)
Source of adapated from D. M. Blei et al. [28].

The representation of the probabilistic model owned by Latent Dirichlet Allocation has three hierarchical levels. [28] generates marginal density calculations in a document that has a probability of acquisition with the equation notation (1).

$$\boxed{p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{Z_{dn}} p\,(Z_{dn}|\theta_d) p(W_{dn}|Z_{dn},\beta) \right) d\theta_d \qquad \textbf{(1)}}$$

BERTopic is one of the topic modeling methods that utilizes the Bidirectional Encoder Representation of the binding transformer. BERTopic allows estimation of the probability of occurrence of a topic in each item based on cluster estimation [29]. Dynamic Topic Modeling is one of the topic models available and provided by BERTopic. Dynamic topic modeling (DTM) is a technique that aims to analyze topics over time. The representation of topic calculations at each time t can be done and adjusted evenly on the representation while maintaining certain words.

2.4. Weighting

In general, weighting uses the term Frequency-Inverse Document Frequency (TF-IDF) to convert words to vectors. The term Frequency TF is used to calculate the number of occurrences of each word (time) and each document. [30]. In contrast, inverse document frequency (IDF) distributes values across. Other terms for the equation differ from the modified BERTopic. The TF-IDF method significantly combines the equation with the notation (2).

$$\boxed{idf(w) = \log\left(\frac{N}{df_i}\right)} \qquad \textbf{(2)}$$

The term frequency is *t* in document *D*, while the document frequency measures the information of a term to a document. Term frequency (TF) is an important concept in information retrieval and text mining, which indicates how often a term *t* appears in a document *D*. It serves as an indicator of the importance of a term in a particular document. However, term frequency alone is not enough to determine the overall significance of a term across a corpus of documents. This is where document frequency (DF) comes into play. Document frequency measures how many documents in the entire corpus contain a term *t*. The intuition behind DF is that a term that appears in many documents is less informative about a particular document. To balance term frequency with document frequency, we use the inverse document frequency (IDF), which is calculated by taking the logarithm of the ratio of the total number of documents to the number of terms. in corpus *N* to the number of documents containing term *t*. The equation for IDF Frequency is calculated by taking the logarithm of the total documents in corpus N divided by the number of documents containing the term with the equation notation (3).

$$\boxed{w_{t,c} = tf_{t,c}\ .\ \log\left(1 + \frac{A}{tf_t}\right)} \qquad \textbf{(3)}$$

The term frequency model is designed to capture the frequency of terms in relation to a specific term *t* within a class c. It essentially combining the set of documents within a particular cluster into single, unified document. This consolidated document represents the collective occurrences of terms, helping to highlight the most relevant terms associated with that cluster. By doing so, the model allows for a more focused understanding of the key terms in each class, which can be used for various tasks such as classification, clustering, and information retrieval.

https://doi.org/10.31849/digitalzone.v15i2. 21030

2.5. Evaluation Matrics

Topic coherence is a widely used evaluation metric for measuring quality of a topic model. It assesses the semantic similarity between the top-ranking words in a topic, helping to determine how interpretable and meaningful the topics are for human understanding. Topic coherence evaluates assesses whether the words that make up a topic are logically and semantically related to one another. It helps determine if the words within a topic form a coherent and meaningful group. Typically, evaluations of topic models rely on multiple coherence metrics to provide a more comprehensive measure of how well the topics align with human understanding and interpretation, ensuring that the topics are both relevant and interpretable. These metrics often focus on the relationships between top words, considering their co-occurrence patterns and semantic similarity coherence measures, including $\boxed{C_{UMass}}$, $\boxed{C_{Uci}}$ and $\boxed{C_V}$, each of which offers a unique perspective on coherence.

UMass coherence relies on document co-occurrence statistics, UCI coherence uses point-by-point mutual information, and Cv coherence combines features from UMass and UCI along with additional vector-based similarity measures. Besides coherence measures, runtime performance is another crucial factor, reflecting the efficiency of a topic model in processing and generating topics. By using a combining of these metrics, researchers and practitioners can gain a thorough understanding of both the quality and efficiency of their topic models. This approach ensuring that the generate topics are not only coherent and meaningful topics but also produced in an efficient manner. One such metric, Unnormalized Coherence Index (UCI) is used assess the quality of topics in a topic model. It focuses on measuring the similarity between words in a topic, ensure that the words are strong related to one another [30]. Equation (5) measures the UMass metric based on documents in a co-occurrence, while the Uci measurement step can be represented by the notation of equation (4).

$$C_{uCi} = \frac{2}{N.(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} PMI(w_i, w_j) \qquad \mathbf{(4)}$$

UMass Coherence is a metric utilized in this study to evaluate topic coherence within the model. It measures how closely related the words in a topic are by examining the distribution of words in documents in a text corpus [30]. UMass Coherence calculates the coherence value by using the frequency of occurrence of words in a document explicitly using the notation equation (5).

$$C_{uMass} = \frac{2}{N.(N-1)} \sum_{i-2}^{N} \sum_{j=1}^{i-1} log + \frac{P(w_i, w_j) + \varepsilon}{P(w_j)} \qquad \mathbf{(5)}$$

The Coherence Matrix value is anothermetrics used to assess the quality of topics produced by a topic model. It evaluates how semantically related the words within a topic are to one other. CV coherence is often used in combination with other coherence metrics to get a more complete picture. Representation of the Cv coherence measurement with equation notation (6).

$$C_v = \left\{ \sum_{w_i \in W`} NPMI\left(w_i, w_j\right)\gamma \right\} J = 1, \ldots., \ldots |W| \qquad \mathbf{(6)}$$

The overall process proposed by us above is one of the measurements used in the research conducted when performing a matrix evaluation related to emerging trends. This approach helps in systematically assessing and analyzing the various patterns and trends within the data. By applying this method, researchers are able to identify significant trends more effectively, allowing for a deeper understanding of the underlying dynamics and providing insights into how these trends evolve over time. The matrix evaluation serves as a crucial tool in organizing and interpreting complex data, ensuring that the analysis is both thorough and insightful.

3. Results and Discussion

The results and discussion section contains the results including The experimental part of this study involves presenting a comprehensive comparison of the well-known methods in the field of topic modeling, namely BERTopic and Latent Dirichlet Allocation (LDA). In this experiment, we apply and test these modeling techniques on the same dataset while maintaining consistent parameter settings across the methods. Our evaluation focuses on two main aspects: a measure of coherence with the architecture and parameters that provide a holistic view of the performance of each method in an optimal manner.

The coherence measures used include UMass, UCI, and Cv, which offer a detailed assessment of the quality of the generated topics. Also before delving into the results of this comparison, we first present the tools and datasets used by the researchers in this study. The dataset was sourced from Twitter using its API, which provides a rich and diverse set of data points relevant to the creative economy and tourism industry. The tools used for preprocessing include tokenization, stemming, and stop word removal to ensure the data is clean and ready for analysis. This rigorous experimental setup ensures that our findings are robust and reliable offering valuable insights into the strengths and weaknesses of each topic modeling method studied.

3.1. Enviroment Tools

In general, the programming language used in our research involves extensive use of Python due to its flexibility and the large number of libraries available for natural language processing and topic modeling. For natural language processing tasks such as tokenization, stopwords removal and stemming, we rely on the Natural Language Toolkit (NLTK) library. NLTK provides a comprehensive suite of tools for working with human language data, making it an essential component of the NLTK preprocessing pipeline [31]. For topic modeling, we leverage Gensim, an open-source library that scales specifically to handle large text corpora and provides complex matrix evaluation [23]. Gensim excels at performing a variety of tasks related to constructing document representations, identifying topics, and analyzing semantic structures in text.

Gensim allows to efficiently transform processed text data into a format suitable for topic modeling and apply sophisticated algorithms to extract meaningful topics from the corpus such as measuring coherence [23]. Interactive development tools facilitate complex code writing, testing, and visualization. Our computing configuration includes a 2.40 GHz CPU with 8 GB of RAM and a 256 GB NVMe disk, which provides sufficient resources to handle the computational demands of our experiments. This configuration allows us to perform model tuning and evaluation to ensure that our topic model is accurate and efficient. Leveraging these sophisticated tools and resources, we can conduct thorough and effective research in the field of topic modeling with a table-based experimental setup (2).

Table 2. Experimental Setup LDA and BERTopic

Models	Experimental Setup
LDA	Preprocessing: Translate Cases , lowercasing, normalization, tokenize, removal stopwords using NLTK, Stemmer using Sastrawi, Count Vectorizer for tokenization, Document-term matrix create Gensim, Model Built using models.lda function number topics from 0 to 9, Coherence score calculated for each optimal number of topic using gensim.
BERTopic	Preprocessing: Translate Cases, lowercasing,

https://doi.org/10.31849/digitalzone.v15i2. 21030

Models	Experimental Setup
	normalization, tokenize, removal stopwords using NLTK, Stemmer using Sastrawi, Model Built using models.BERTopic number topics 10 from, Count Vectorizer for tokenization, c-TF-IDF Coherence score calculated for each optimal number of topic using gensim.

Our research offers an architecture represented by custom embeddings sub models from BERT, in our discussion we also do this by combining using clustering available in BERT. We provide default parameters to provide cluster performance on the sub model algorithm as an optimal option. Completely the parameters in the topic of use that we propose are based on the architecture in the research that we produce in the representation in the table (3).

Table 3. Architecture and hyperparameter model topic

Models	Algorithm and experimental setup
LDA	Corpus set from gensim_corpus create, id2word from dictionary vector, num_topics set 10 topics, chunksize=100, passes=50, iterations default 50, alpha and eta set using symmetric, per_word_topics set boolean true.
BERTopic	Corpus set from gensim corpus and dictinonary vector, embeddings custome using hugging multilingual-MiniLM-L12-v2, Clustering HDBSCAN without UMAP with min cluster size=10, sample=5, matric using Euclidean cluster method eom and prediction set for true, based architect parameter BERT vebose set boolean true, calculated for boolean true.

The number of topics in the topic latent must be selected before LDA and BERTopic, we adjust both topic latent models with a parameter value of 10 for the specified number of latent topics. Process the above parameters to optimize [31] latent. Alpha and beta are the Dirichlet prior values in LDA [31] we apply priors with symmetric parameters in the two hyperparameters that we apply to the table (3) Another process influence in the parameters in our memory consumption power optimization is set in the chunksize amount.

Implement BERTopic using custom embedding from hugging, our steps from embedding without using UMAP aim to explode the memory usage of embedding without setting the low_memory parameter to regulate its prevention. [31]. In addition, we also apply verbose to true values as a step in tracking the stages of the model [31] while the sum value latent we set the same as the LDA value in the trend comparison.

3.2. Preprocessing Dataset

The data source used in the research to conduct the experiment was a collection of tweets taken by the researcher every month based on a time span each year using relevant keywords in table (1) by the researcher. Tweet data retrieval with binding API [31] and keywords and usage provides researchers with the flexibility to retrieve source data for each tweet about tourism, tourism, and creative economy. Data retrieval obtained from tweets provides unstructured data which is a challenge in itself in making topic modeling more challenging. The dataset includes 65,806 tweet. The results of preprocessing and crawling of each tweet are placed in its own text file.

Data preprocessing is very important and affects the results of the experiment substantially. Data must be pre-processed to ensure consistency in experiment format [21], we did text preprocessing and some changes to the dataset. We plan to conduct research to measure the effects of stemming and stopwords on LDA and BERTopic models. In the first step, we proceed to lowercase, and the removal of any words consisting of one character tokenization, stopword removal, stemming. The process that has been done, is continued to represent the term vector in the data size including vocabulary and the vector size represented in table (4).

Table 4. Ablation Data Compared Processing

Models	Descriptions	Vocabulary	Term id Vector
LDA	With Stopwords and Stemming	67725	52962
BERTopic	With Stopwords and Stemming	67725	52962

Our data describes in the model adjustment of the overall steps before the process continues with parameter retrieval and matrix evaluation. In this case we try to normalize the vocabulary according to the vocabulary provisions causing a reduction in rare words in the document. This step is intended to try to provide optimal steps in our preprocessing It is very important to observe the data used, paying attention to the number of vocabulary and vector ids greatly influences the vector results that will be used in measuring the matrix.

3.3. Evaluation Experiments of Research

Experiment and results of data processing, implementation, and verification of BERTopic and LDA models. In this study, we aim to compare the evaluation metrics of both BERTopic and LDA topic modeling using UCI, Cv, UMass topic coherence. At the same time, we aim to evaluate the impact of stopword and stemming tasks on topic modeling methods. The considered models and their corresponding vocabulary sizes are illustrated in Table (4). The same technique as our method tries to do on the effects of a fairly optimal model produces the evaluation values of the coherence matrix represented in the table (5).

Table 5. Coherence Score for topic models

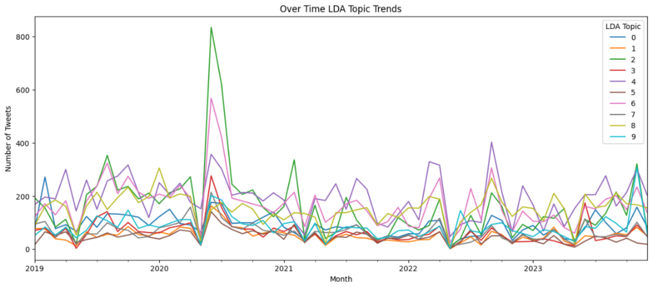
Models	Score c_uci	Score c_v	Score u_mass
BERTopic	0.16	0.53	-6.252
LDA	-3.163	0.30	-7.305

The evaluation data in table (5) provides insight that the model measurement is quite optimal where the U-mass matrix evaluation is less than optimal due to the availability of vocabulary data in each document which is minimal with the availability of words in sentences giving a negative effect on the measurement of the coherence value. Topic terms are able to provide a level of representation in labeling and naming a topic [21]. Temporally changing values are metrics used to show changes over time in a particular topic or term [21]. The amount of data that is preprocessed gives a minimal frequency of occurrence of words in a document which also makes it difficult for the model to find topics.

The high sparsity in the corpus has a significant negative impact on the coherence evaluation. The negative impact of the coherence value representation can be seen using the topic label which is also used in the evaluation, where the topic label represents the topic results over time. Trends that change over time can be repeated and insight into trend knowledge of a topic label. The over time representation of the topic label in the model is known in table (6).

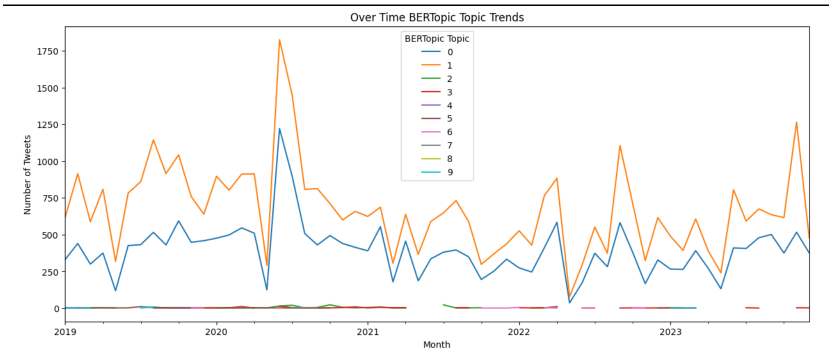
Table 6. Over Time Topic Model Comparasion

Over Time LDA Trends



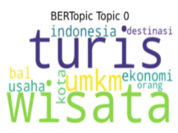
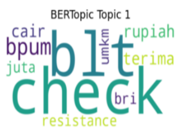
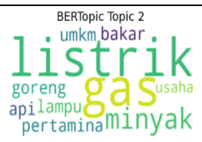
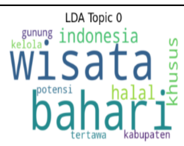
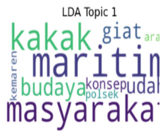
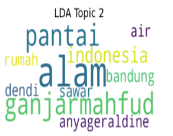
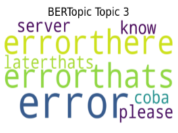
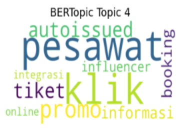
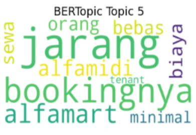
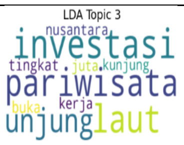
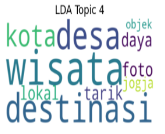
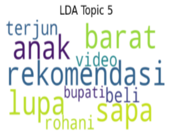
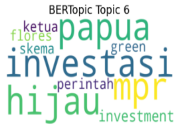
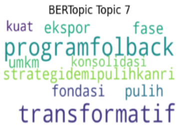
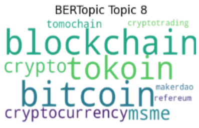
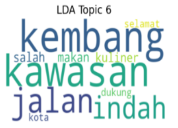
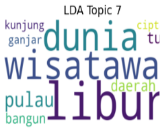
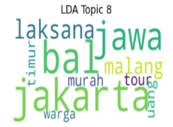
Over Time BERTopic Trends

https://doi.org/10.31849/digitalzone.v15i2. 21030



Visualization of the resulting trend shows fluctuations in topic label results that change and experience a fairly high spike in mid-2020 to 2021. The visualization still does not ignore outlier topics from each model to produce transparent experiments. The topic label results produced by BERTopic are biased and more dominant on label 1, where label 1 produces the most topics over time. Meanwhile, latent Dirichlet allocation produces labels that are quite ideal in the distribution of the resulting labels. In the insight of each trend, it is necessary to know the identification of the topics produced. This insight can be represented in the table (7) as well as identifying topics whether there are differences in each topic produced by the model, topic identification from this study looks at the topic results in more detail by looking at the words produced. Word cloud is able to create something that can reflect the features of stage words [26]. The identification results will only represent a number of topics that are closest to the topic as a whole.

Table 7. Topics Identified

Toursim and Local Economy	**Financial assistance and ricect aid**	**Energy Consumption**	**Tourism and Natural Attractions**	**Community and Activites**	**Beach and Natural Scenery**
					
Technical Support	**Travel Online Booking**	**Rental and Service Cost**	**Investment and Tourism Growth**	**Destinations and Attractions**	**Recreational and Personal Recomend**
					
Green Investment and Government	**Growth Strategi Planning**	**Cryptocurrency and Blockchain Market**	**Culinary and Scenic Areas**	**Internasional Toursim**	**Regional Travel and Economy**
					

Evaluation of identification needs to be done, the results of the two topics have differences in the resulting topic labels, where there are differences related to government, technical support and cryptocurrency blockchain market, while when analyzed against the results of the identification there are similarities regarding tourism, online travel and investment. then follow-up in the analysis results is carried out to see more details on the 3 topics that have closeness or have similar topics between topics. The evaluation results are designed in a table evaluation (8) with the aim of seeing the similarity of topics between topics. From the results of the experimental trials, the researcher determined the three topics that had the highest level of similarity.

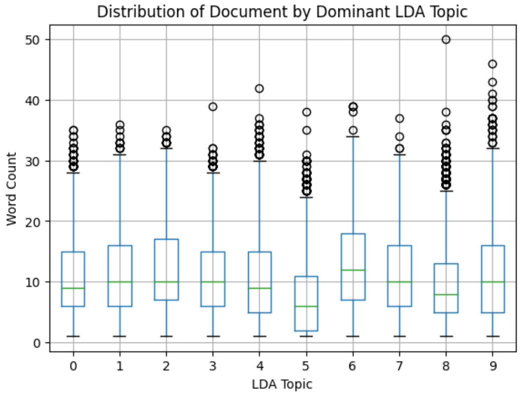
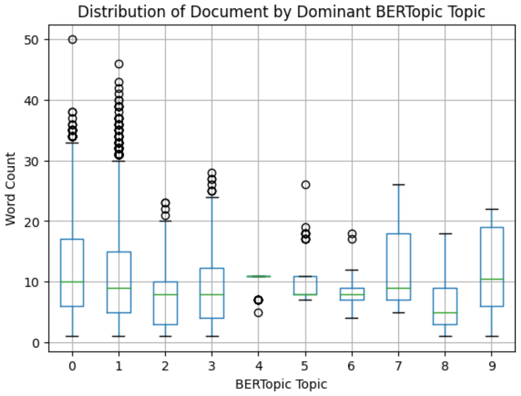
The results of the comparison in the topic grouping had the highest similarity between topics 0 and 1 in the BERTopic similarity. While the similarity produced by LDA had the highest level of similarity on topics 0 and 4. Each topic produced the lowest similarity value of 0.0049 and 0.0002 on topics four and five. Overall, the details of the similarity can be represented in the table (8). Dimension reduction can help to project a data value into a dimension [22]. The low dimensionality of the data values causes the resulting clusters to be able to have a sensitive impact on the model. To handle this, it is necessary to ensure tokenization of the HDBSCAN cluster in the corpus calculation to increase the uniqueness and specificity of the resulting topics [22].

Table 8. Most similarity topics for topics

Topic	Cosine Similarity BERTopic	Topic	Cosine Similarity LDA
T0	Topic 1 (0.9477), Topic 3 (0.3037), Topic 6 (0.2063)	T0	Topic 4 (0.8745), Topic 2 (0.0386), Topic 5 (0.0002)
T1	Topic 0 (0.9477), Topic 3 (0.2694), Topic 6 (0.1990)	T1	Topic 8 (0.0004), Topic 5 (0.0004), Topic 3 (0.0003)
T2	Topic 0 (0.2032), Topic 1 (0.1899), Topic 3 (0.0909)	T2	Topic 0 (0.0386), Topic 3 (0.1333), Topic 7 (0.0951)
T3	Topic 0 (0.3037), Topic 1 (0.2694), Topic 2 (0.0909)	T3	Topic 7 (0.0308), Topic 8 (0.0170), Topic 2 (0.0105)
T4	Topic 0 (0.0108), Topic 1 (0.0079), Topic 2 (0.0049)	T4	Topic 0 (0.8745), Topic 6 (0.0104), Topic 5 (0.0001)
T5	Topic 1 (0.1102), Topic 0 (0.0964), Topic 2 (0.0259)	T5	Topic 6 (0.0004), Topic 8 (0.0004), Topic 3 (0.0004)
T6	Topic 0 (0.2063), Topic 1 (0.1990), Topic 2 (0.0542)	T6	Topic 7 (0.0118), Topic 4 (0.0104), Topic 9 (0.0078)
T7	Topic 1 (0.1601), Topic 0 (0.1498), Topic 9 (0.0840)	T7	Topic 3 (0.0308), Topic 8 (0.0167), Topic 6 (0.0118)
T8	Topic 0 (0.1392), Topic 1 (0.1135), Topic 2 (0.0679)	T8	Topic 3 (0.0170), Topic 7 (0.0167), Topic 5 (0.0004)
T9	Topic 0 (0.1250), Topic 1 (0.1078), Topic 7 (0.0840)	T9	Topic 6 (0.0078), Topic 5 (0.0003), Topic 8 (0.0002)

The shift in the topic is represented in the distribution table. Distribution plays an important role in grouping in this study. Distribution insights that affect the dominance of each topic are able to provide a sensitive impact on the topic results. Where it can be seen that BERTopic almost entirely dominates words into topics 0 and 1, dropping drastically in the distribution of topic results 4. The fluctuating shift has a negative impact even though it has a fairly high coherence value in its distribution. The difference is very visible in the LDA model which is able to distribute evenly from the model results. The results of both distributions provide insight for further research.

Table 9. Distribution of document

LDA	BERTopic
	

Social media that plays an important role in conveying information can provide insight into systematic identification and theme analysis. A structured approach to the results of the experiment provides information that supports recommendations such as accommodation of transportation services for tourists, government services that are able to provide innovation for

https://doi.org/10.31849/digitalzone.v15i2. 21030

incoming investors or investment. The resulting topic proposals provide a comprehensive framework for analysis by utilizing topic modeling insights.

The proposed performance provides an optimal step in topic distribution. The proposed theme performance in visualization provides a view of the relationship between word clouds, similarity matrices, and topic evaluations over time in evaluating trends in the creative economy and tourism. Insights that provide topic modeling value can also help in optimizing the strategic planning framework for governments such as innovation and investment. Encouraging such as management aspect policies, target processes that support aspects of the creative economy and tourism towards tourism. The results of the theme also provide information that there needs to be a level of effort on areas that need to be improved in exploration using social media as topic information.

4. Conclusion

Evaluation of the creative economy and tourism in the objectives of researchers who focus on social media, the study provides a comparison of the results of two popular techniques and are often used by previous studies, the proposed experimental setup such as hyperparameters and architecture to the preprocessing setup of the data provides quite good computation. The results of the study based on coherence, distribution, similarity, overtime. BERTopic provides topics that are able to categorize hidden topics such as cryptocurrency bitcoin to energy consumption, the value of the matrix size is measured from the coherence value which is categorized as good, while LDA is able to distribute topics in topic stability. The high relevance of the LDA results is measured in the evaluation of the matrix which is categorized as sufficient.

Topics for five years provide an overview as a breakthrough step related to tourists and transportation accommodation. Where transportation services in the theme of creative economy and tourism provide a trend that is always explored for social media users. The relationship that has an impact on this becomes the main domain. In identifying a fairly good distribution between the two topics, it gives a different latency. BERTopic gives a very different topic latency from LDA so that it is necessary to do in terms of deepening the preprocessing data which also needs to be improved and further investigation.

Optimization of the trial of using different parameters is able to provide high coherence values, vectorization and the number of vocabulary determines the value of the matrix. This is a further challenge and needs to be improved in the process for preprocessing Indonesian. Seeing the increase in regional languages that are often used to communicate on social media, provides innovation to topics that can always change in the resulting model grouping. The decline in topic results can also occur due to the emergence of words that do not provide meaning so that they are unable to provide relevance to other words. Thus clarifying further exploration to provide research insights that are able to provide identification so that they can be applied in overcoming the creation of creative economic and tourism policies or informing the relevance process related to the topics identified into decision making to encourage optimistic results as a future economic sector.

References

[1] M. Ashoer *et al.*, *Ekonomi Pariwisata*, vol. 5, no. 3. 2021.

[2] A. S. E. S. Sri Hardianti Sartikan, Muhammad hasan, *Ekonomi Kreatif Agus Syam*, no. July. 2022.

[3] KEMENPAREKRAF, “Outlook Pariwisata dan Ekonomi Kreatif,” *Deputi Bid. Kebijak. Strateg. Kementeri. Pariwisata dan Ekon. Kreat. Badan Pariwisata dan Ekon. Kreat. Republik Indones. Jakarta – Indones.*, pp. 1–68, 2020, [Online]. Available: https://bankdata.kemenparekraf.go.id/upload/document_satker/a6d2d69c8056a29657be2b5ac3107797.pdf.

[4] Kementrian Pariwisata dan Ekonomi Kreatif, “Wisatawan Mancanegara Perkembangan Januari 2024,” pp. 4–4, 2024.

[5] We Are Social Meltwater, "Digital 2023 Report," *Meltwater*, p. 213, 2023.

[6] A. Yunita, S. F. Telaumbanua, and A. Irawan, "The Tweetology of New and Renewable Energy in Indonesia," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 17, no. 2, p. 127, 2023, https://doi.org/10.22146/ijccs.81397.

[7] F. Gurcan, A. Ayaz, G. G. Menekse Dalveren, and M. Derawi, "Business Intelligence Strategies, Best Practices, and Latest Trends: Analysis of Scientometric Data from 2003 to 2023 Using Machine Learning," *Sustain.*, vol. 15, no. 13, 2023, https://doi.org/10.3390/su15139854.

[8] W. Kang, Y. Kim, H. Kim, and J. Lee, "An Analysis of Research Trends on Language Model Using BERTopic," *Proc. - 2023 Congr. Comput. Sci. Comput. Eng. Appl. Comput. CSCE 2023*, no. 2022, pp. 168–172, 2023, https://doi.org/10.1109/CSCE60160.2023.00032.

[9] L. Yijia, "Comparison of LDA and BERTopic in News Topic Modeling : A Case Study of The New York Times ' Reports on China," vol. 7, no. June, pp. 47–51, 2024, https://doi.org/10.55014/pij.v7i3.616.

[10] F. Bianchi, S. Terragni, and D. Hovy, "Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence," *ACL-IJCNLP 2021 - 59th Annu. Meet. Assoc. Comput. Linguist. 11th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, vol. 2, pp. 759–766, 2021, https://doi.org/10.18653/v1/2021.acl-short.96.

[11] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Eng. Appl. Artif. Intell.*, vol. 110, no. February, p. 104743, 2022, https://doi.org/10.1016/j.engappai.2022.104743.

[12] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Naacl-Hlt 2019*, no. Mlm, pp. 4171–4186, 2018, [Online]. Available: https://aclanthology.org/N19-1423.pdf.

[13] C. D. P. Laureate, W. Buntine, and H. Linger, *A systematic review of the use of topic models for short text social media analysis*, vol. 56, no. 12. Springer Netherlands, 2023.

[14] Y. Sahria and D. Hatta Fudholi, "Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode Topic Modeling LDA (Latent Dirichlet Allocation)," *Masa Berlaku Mulai*, vol. 1, no. 3, pp. 336–344, 2017.

[15] A. Panichella, "A Systematic Comparison of search-Based approaches for LDA hyperparameter tuning," *Inf. Softw. Technol.*, vol. 130, p. 106411, 2021, https://doi.org/10.1016/j.infsof.2020.106411.

[16] S. J. Blair, Y. Bi, and M. D. Mulvenna, "Aggregated topic models for increasing social media topic coherence," *Appl. Intell.*, vol. 50, no. 1, pp. 138–156, 2020, htpps://doi.org/10.1007/s10489-019-01438-z.

[17] M. M. Mostafa, *A one-hundred-year structural topic modeling analysis of the knowledge structure of international management research*, vol. 57, no. 4. Springer Netherlands, 2023.

[18] R. Panchendrarajan and A. Saxena, "Topic-based influential user detection: a survey," *Appl. Intell.*, vol. 53, no. 5, pp. 5998–6024, 2023, doi: 10.1007/s10489-022-03831-7.

[19] S. Hwang, E. Flavin, and J. E. Lee, "Exploring research trends of technology use in mathematics education: A scoping review using topic modeling," *Educ. Inf. Technol.*, vol. 28, no. 8, pp. 10753–10780, 2023, doi: 10.1007/s10639-023-11603-0.

[20] Z. A. Hasibuan, "Towards using universal big data in artificial intelligence research and development to gain meaningful insights and automation systems," *2020 Int. Work. Big Data Inf. Secur. IWBIS 2020*, pp. 9–15, 2020, https://doi.org/10.1109/IWBIS50925.2020.9255497.

[21] O. Ozyurt, H. Özköse, and A. Ayaz, "Evaluating the latest trends of Industry 4.0 based on LDA topic model," *J. Supercomput.*, no. 0123456789, 2024,

https://doi.org/10.1007/s11227-024-06247-x.

[22] M. Mishra, S. K. Vishwakarma, L. Malviya, and S. Anjana, "Temporal analysis of computational economics: a topic modeling approach," *Int. J. Data Sci. Anal.*, 2024, https://doi.org/10.1007/s41060-024-00596-9.

[23] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, "Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study," *Electron.*, vol. 12, no. 12, 2023, https://doi.org/10.3390/electronics12122605.

[24] C. Lalk *et al.*, "Measuring Alliance and Symptom Severity in Psychotherapy Transcripts Using Bert Topic Modeling," *Adm. Policy Ment. Heal. Ment. Heal. Serv. Res.*, vol. 51, no. 4, pp. 509–524, 2024, doi: 10.1007/s10488-024-01356-4.

[25] I. Lauriola, A. Lavelli, and F. Aiolli, "An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools," *Neurocomputing*, vol. 470, no. xxxx, pp. 443–456, 2022, doi: 10.1016/j.neucom.2021.05.103.

[26] B. Yin and C. H. Yuan, "Detecting latent topics and trends in blended learning using LDA topic modeling," *Educ. Inf. Technol.*, vol. 27, no. 9, pp. 12689–12712, 2022, htpps://doi.org/10.1007/s10639-022-11118-0.

[27] Y. Chen, Z. Xie, and D. K. W. Chiu, "Analytics of motivational factors of educational video games: LDA topic modeling and the 6 C's learning motivation model," *Educ. Inf. Technol.*, 2024, https://doi.org/10.1007/s10639-024-12726-8.

[28] D. M. Blei, A. Y. Ng, and M. T. Jordan, "Latent dirichlet allocation," *Adv. Neural Inf. Process. Syst.*, no. January 2001, 2002.

[29] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2020.

[30] A. Farea, S. Tripathi, G. Glazko, and F. Emmert-Streib, "Investigating the optimal number of topics by advanced text-mining techniques: Sustainable energy research," *Eng. Appl. Artif. Intell.*, vol. 136, no. PA, p. 108877, 2024, https://doi.org/10.1016/j.engappai.2024.108877.

[31] H. Zankadi, A. Idrissi, N. Daoudi, and I. Hilal, "Identifying learners' topical interests from social media content to enrich their course preferences in MOOCs using topic modeling and NLP techniques," *Educ. Inf. Technol.*, vol. 28, no. 5, pp. 5567–5584, 2023, https://doi.org/10.1007/s10639-022-11373-1.