Improving Spatio-Temporal Topic Modeling with Swarm Intelligence: A Study on TripAdvisor Forum of Morocco

Original Scientific Paper

Ibrahim Bouabdallaoui*
LASTIMI Laboratory – High School of Technology Salé, Mohammed V University in Rabat
Avenue Le Prince Héritier, Salé, Morocco
ibrahim_bouabdallaoui@um5.ac.ma

Fatima Guerouate
LASTIMI Laboratory – High School of Technology Salé, Mohammed V University in Rabat
Avenue Le Prince Héritier, Salé, Morocco
fatima.guerouate@est.um5.ac.ma

Mohammed Sbihi
LASTIMI Laboratory – High School of Technology Salé, Mohammed V University in Rabat
Avenue Le Prince Héritier, Salé, Morocco
mohammed.sbihi@est.um5.ac.ma

*Corresponding author

Abstract *– This study introduces innovative methodologies for spatiotemporal topic modeling applied to the TripAdvisor forum of Morocco, leveraging the diverse and geographically tagged user-generated content. We develop and evaluate two schemas integrating Latent Dirichlet Allocation (LDA) with advanced clustering techniques, including a hybrid K-Means algorithm that incorporates Genetic Algorithms and the Artificial Bee Colony method. The first schema independently processes user threads, publication times, and locations using LDA, followed by clustering, while the second schema combines these dimensions into a unified vector for holistic LDA application, facilitating direct comparisons of clustering efficacy. Our findings demonstrate that swarm intelligence significantly boosts clustering performance, especially for larger clusters, and enhances the visualization of complex data relationships. These insights offer actionable intelligence for tourism stakeholders and underscore the practical benefits of advanced computational techniques in harnessing user-generated content for strategic decision-making.*

Keywords: *topic modeling, latent Dirichlet allocation, artificial bee colony, genetic algorithms, k-means*

Received: March 24, 2024; Received in revised form: June 14, 2024; Accepted: June 17, 2024

1. INTRODUCTION

The digital era has catalyzed an unprecedented expansion of online content, transforming forums into invaluable repositories of user-generated data. Among these, TripAdvisor stands out as a premier global travel platform, amassing a vast array of reviews, discussions, and user interactions that are rich in spatial and temporal diversity. This platform provides a unique window into user experiences, offering insights across a spectrum from travel advice to detailed service reviews [1]. The Moroccan TripAdvisor forum, in particular, encapsulates a vivid tableau of the region's cultural, economic, and touristic pulse. However, the complexity and volume of this data pose significant analytical challenges, underscoring the necessity for sophisticated analytical methodologies. In the realm of text analysis, Topic Modeling, particularly through the use of Latent Dirichlet Allocation (LDA), has proven to be a powerful tool for uncovering latent thematic patterns within large text corpora. The integration of spatial and temporal data into topic modeling further enhances our ability to perform dynamic, context-aware analyses—aptly termed Spatio-Temporal Topic Modeling. This approach finds applicability in a myriad of fields such as urban planning, epidemiology, and notably, tourism, where understanding spatial and temporal variations is crucial. Recent studies underscore the growing sophis-

tication in spatio-temporal topic modeling across various domains, affirming the relevance of our proposed methodologies for the TripAdvisor forum of Morocco. Liu et al. (2015) demonstrated the potential of spatio-temporal topic models to analyze social media check-in data, revealing user movements and interests that parallel the tourist behaviors observable in TripAdvisor reviews [2]. Similarly, Min et al. (2014) explored multimodal spatiotemporal themes in landmark studies, a concept that can be adapted to identify and analyze thematic patterns in reviews related to specific tourist landmarks [3]. Luna and Genton (2005) offered a predictive model approach for handling spatially sparse but temporally rich data, an approach that could enhance the understanding of spatial and temporal dynamics in TripAdvisor forum data [4]. Additionally, the work by Chen et al. (2019) on local topic detection using spatio-temporal social media data provides a valuable framework for extracting localized insights from geographically tagged discussions on TripAdvisor [5]. Finally, the methodology proposed by Zhao et al. (2016) for efficiently mining topics from spatio-temporal documents could directly inform our approaches to managing the complex dataset derived from the TripAdvisor forum [6]. Collectively, these studies not only validate the necessity of advanced modeling techniques but also enhance the robustness of our research design, aiming to uncover rich, actionable insights into the tourism dynamics depicted in user-generated content. This study introduces two innovative schemas for applying Spatio-Temporal Topic Modeling to the TripAdvisor forum of Morocco. These schemas aim to synergistically combine text, temporal, and spatial data using advanced methodologies including LDA, vectorization, autoencoding, and a novel hybrid K-Means clustering approach that integrates the capabilities of Genetic Algorithms and the Artificial Bee Colony method. The objective is to evaluate these schemas' effectiveness in generating discernible, meaningful topics and improving clustering performance for large, complex datasets.

2. LITERATURE REVIEW

This paper builds upon a rich body of work in the field of spatio-temporal Topic Modeling. It is, therefore, imperative to discuss and understand the relevant research landscape. This section presents a review of the pertinent literature, tracing the development of key concepts, identifying the primary methodologies employed, and highlighting significant findings and their implications. It also identifies gaps in the existing research, underscoring the contribution of our study to the field of spatio-temporal Topic Modeling. Researchers introduced a comprehensive framework for managing, processing, analyzing, and detecting trending topics in streaming data coming from Twitter [7]. Their utilization of a hybrid model selector and their application of deep learning and transfer learning techniques for classifying health-related tweets are noteworthy. The paper presents a methodical approach to topic detection, focused on processing data with sentence granularity, pertinent to the nature of Twitter messages. It engages a variety of techniques including Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), LDA-MALLET, and Biterm Topic Modeling to assess the effectiveness of the proposed framework. These models are used for data dimensionality reduction, clustering of documents, and in-depth analysis of short messages, respectively. Our work parallels this research, particularly in its emphasis on the spatio-temporal aspects of topic modeling. However, our approach is uniquely applied to the domain of tourism, focusing on user-generated content from TripAdvisor in the context of Morocco, rather than health-related Twitter data. Like the study above, our research also grapples with the complexities of conflicting data from different locations and times. We also share a commitment to cleaning and preprocessing data to enhance quality. In addition, both studies underline the importance of visualization to understand the topics' dynamics. A paper presents an innovative two-stage system to detect and track events from tweets [8]. By integrating Latent Dirichlet Allocation (LDA) and a density-contour-based spatio-temporal clustering approach, it manages to create a comprehensive framework for tracking events on Twitter, where events are identified as topics in tweets. The event identification process involves partitioning the geo-tagged tweet stream into temporal windows and running an LDA-based topic discovery step. Subsequently, each tweet is assigned an event label, and density-contour-based spatio-temporal clustering is employed to identify event clusters. Our study resonates with this approach as we also employ LDA for topic modeling in our Schema I. However, in Schema II, we diverge by using swarm intelligence algorithms to form a combined context content vector, hence tracking the dynamic spatiotemporal trends in a more nuanced way. The paper also introduces a novel methodology for ensuring topic continuity through calculating KL-divergences between topics, and a density-contour clustering approach for establishing spatio-temporal continuity. These methodologies bear similarities with our temporal and spatial correlation strategies in the dynamic clustering process. Their work, like ours, emphasizes the significance of spatio-temporal analysis in understanding the dynamics of topic trends, thereby contributing to the broader field of spatio-temporal topic modeling. However, their focus on tracking events in Twitter using density-contour-based clustering, while ours focuses on swarm intelligence, highlights different approaches to the similar challenge of tracking dynamic trends. A novel approach to topic modeling has been presented by authors who developed a multi-objective optimization algorithm based on the swarm intelligence of a bee colony, known as the Multi-Objective Artificial Bee Colony (MOABC) [9]. This method aims to enhance the performance of topic modeling, an area of text analysis that extracts underlying topics from document collections. Traditionally,

Latent Dirichlet Allocation (LDA) has been the most recognized method for topic modeling. LDA models each document as a probabilistic distribution over latent topics, considering a multinomial distribution for the document and the topics, each generated from a Dirichlet distribution with specific parameters. However, the authors argue that there is room for significant improvements in LDA's performance. To address this, the MOABC algorithm has been introduced. The methodology of the MOABC algorithm incorporates several steps: initializing the set of non-dominated solutions, assigning solutions to each bee, executing the main loop of the algorithm, generating modified solutions, replacing original solutions if the modified ones are better, sorting solutions by ranking and crowding, and updating the set of non-dominated solutions until a maximum number of cycles are reached. This approach essentially considers multiple objectives such as coherence, coverage, and perplexity, and each solution represents a set of topics along with their respective weights. The results from the experiments conducted on the Reuters-21578 and TagMyNews datasets indicated that the MOABC approach provides relevant improvements with respect to both LDA and the Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [10]. Therefore, the study provided substantial evidence supporting the exploitation of the multi-criteria nature of topic modeling with multi-objective optimization approaches, which marks an important development in the field of topic modeling. A review highlights the applicability of transformers for modeling long-range dependencies across various domains, including NLP [11]. The paper discusses how transformers, which have been successful in NLP, can be adapted for spatio-temporal modeling in different modalities, offering insights that could be applied to spatio-temporal topic modeling in textual data analysis. A work extends traditional language modeling to include spatiotemporal conditions, providing a novel approach for modeling text associated with specific times and places. It aims to capture the neighborhood, periodicity, and hierarchy within spatio-temporal text data, offering insights that are directly applicable to understanding and modeling the dynamics of user-generated content on platforms like TripAdvisor, particularly in how text is generated in response to spatial and temporal contexts. The study develops neural network models for language modeling conditioned on spatio-temporal variables [12].

3. METHOD

3.1. DATASET

The dataset used for this study is a result of web scraping from the TripAdvisor forum in Morocco. It consists of a total of 29,733 posts, spanning a significant period of more than 15 years from December 2007 to March 2023. Such a wide time frame presents an excellent opportunity for longitudinal study of trends, and the transformation of topics over time, thereby adding a temporal layer to our analysis. Each entry in the dataset includes the post content, the username of the post's author, their location, and the date of the post. The post content is used as the main body of text for topic modeling, and it provides a rich source of diverse perspectives and experiences shared by the forum users. The username of the post's author adds an element of personalization, potentially allowing for the exploration of user-specific topics or trends. The location information of the authors brings a unique perspective to our study. It allows us to understand the geographical distribution of the authors and to assess any potential influences of the authors' home locations on the topics discussed, thereby adding a spatial dimension to our analysis. The date of each post, indicating when the content was shared, is key for the temporal aspect in LDA. This information helps us understand how topics and discussions have evolved over time and could reveal temporal trends or patterns in the data. Collectively, this dataset offers a wealth of information for conducting a comprehensive spatio-temporal topic modeling study. Its size and depth make it suitable for testing and validating our proposed schemas, while its diversity ensures that the findings of our research are representative and applicable to a wide range of situations in the realm of tourism [13].

3.2. A MULTI-LAYERED APPLICATION OF LDA AND SWARM-ENHANCED CLUSTERING

In the first schema of our methodology (Fig. 1.), we approach the spatio-temporal topic modeling task through a sequential and layered implementation of Latent Dirichlet Allocation (LDA), followed by the independent application of two different swarm intelligence algorithms for clustering.

Stage 1: LDA Implementation:

- Textual LDA: The first stage of this schema involves applying LDA to user threads, with each thread treated as a separate document. By applying LDA, we extract the underlying topics or themes from the text content, each represented by a set of topic probabilities. This allows us to understand the main themes and topics discussed in the user threads, providing a detailed thematic analysis of the text data.
- Temporal LDA: Next, we apply LDA to the time of publication of each thread, treating each timestamp as a temporal document. This allows us to uncover temporal themes, potentially revealing patterns or trends in discussions over time. By analyzing the distribution of topics over time, we can identify how certain topics gain or lose prominence, reflecting temporal dynamics in user discussions.
- Spatial LDA: The third application of LDA is performed on the author's location data associated with each post. This step uncovers spatial themes, indicative of geographically influenced discussions

or trends. By examining how topics vary across different geographical locations, we can gain insights into region-specific interests and trends.

Each of these applications of LDA produces a topic distribution for each document (thread). These distributions are then vectorized to create a unified representation of the textual, temporal, and spatial themes for each thread. This unified representation is crucial for capturing the multifaceted nature of the data, integrating text, time, and location into a comprehensive feature set [14].

Stage 2: Autoencoder Processing

The resulting vectors from the LDA stages serve as input to an autoencoder. Autoencoders are neural networks trained to recreate their input data, thereby learning compressed, meaningful representations of the input data in their hidden layers. The autoencoder learns to encode the high-dimensional input vectors into a lower-dimensional space, capturing the most salient features of the data. This compressed representation is then decoded back to reconstruct the original input, ensuring that the encoded features retain the essential information from the original vectors [14].

Stage 3: Swarm-Enhanced Clustering

Finally, the output from the autoencoder is subjected to clustering. We independently apply two different swarm intelligence algorithms—Genetic Algorithms [15] and Artificial Bee Colony [16] —to benchmark their performance in identifying and forming distinct clusters. These algorithms enhance the capability of traditional K-Means clustering by leveraging their respective exploratory capabilities.

- Genetic Algorithm (GA): This algorithm optimizes the clustering process by iteratively improving the cluster centroids based on selection, mutation, and crossover processes. The GA starts by vectorizing topics and defining K-Means and GA parameters. It then generates a random initial population and iteratively improves it based on the silhouette coefficient until the optimal population is found. This evolutionary approach ensures robust exploration of the solution space, enhancing the clustering quality [15].
- Artificial Bee Colony (ABC): This algorithm mimics the foraging behavior of bees to find the optimal clustering by exploring and exploiting the solution space effectively. The ABC initializes by vectorizing topics and defining ABC setup parameters. It generates a random population and updates employed onlooker, and scout bees iteratively based on the silhouette coefficient until the optimal population is reached. This bio-inspired approach balances exploration and exploitation, leading to effective clustering outcomes [16].

By integrating LDA with swarm intelligence algorithms, our schema (Fig. 1.) represents an integrated, layered approach to topic modeling. Text, time, and location are independently analyzed but ultimately unified to inform a clustering process optimized through the independent application of swarm intelligence algorithms. This comprehensive approach not only captures the multi-dimensional nature of the data but also leverages advanced clustering techniques to produce meaningful and well-defined topic clusters.

3.3. A UNIFIED LDA APPROACH WITH BENCHMARKING OF SWARM-ENHANCED CLUSTERING TECHNIQUES

In the second schema of our methodology (Fig. 2), we address the spatio-temporal topic modeling task through a unified implementation of Latent Dirichlet Allocation (LDA), followed by the independent application of two different swarm intelligence algorithms for clustering. Unlike the first schema, where text, time, and location were independently analyzed, the second schema commences with the concatenation of these elements into a single vector for each document (thread). Each element (text of the post, location of the user, and time of publication) is treated as part of a unified, comprehensive document. This approach allows us to maintain the contextual linkages between these elements, fostering an integrated analysis that inherently reflects their interconnections [17].

Stage 1: Data Concatenation

Each element—text of the post, location of the user, and time of publication—is treated as part of a unified, comprehensive document. This approach allows us to maintain the contextual linkages between these elements, fostering an integrated analysis that inherently reflects their interconnections. By concatenating the text posts, timestamps, and location data into a single vector, we create a multifaceted representation of each document.

Stage 2: LDA on Concatenated Features

Upon constructing these comprehensive vectors, LDA is applied to extract topics. Given the incorporation of textual, temporal, and spatial elements in each vector, the derived topics are inherently spatio-temporal, reflecting themes that capture the interplay between the content of discussions, when they occurred, and where the participants were located. This unified application of LDA allows for a holistic analysis, where the interconnectedness of the different data aspects is preserved and leveraged.

Stage 3: Autoencoder Processing

The resulting topic distributions from the LDA stage serve as input to an autoencoder. Autoencoders are neural networks designed to learn efficient codings of the input data, providing compressed, meaningful representations that retain essential information. This step further refines the representation of the spatio-temporal topics, ensuring that the most salient features are captured and utilized in the subsequent clustering process.

Stage 4: Swarm-Enhanced Clustering

After topic extraction and autoencoder processing, we employ the K-Means clustering algorithm to group these topics. To assess the performance of different optimization strategies for this clustering process, we independently apply two swarm intelligence algorithms—Genetic Algorithms [15] and Artificial Bee Colony [16]. Each of these algorithms is benchmarked against standard K-Means clustering, evaluating their respective capabilities to form distinct and meaningful clusters.

By unifying the analysis of text, time, and location, and benchmarking different swarm-enhanced clustering techniques, the second schema (Schema II) offers an integrated approach to spatio-temporal topic modeling.

It provides valuable comparative insights into the effectiveness of different clustering strategies, demonstrating the benefits of combining LDA with advanced swarm intelligence algorithms for comprehensive and meaningful topic extraction and clustering.

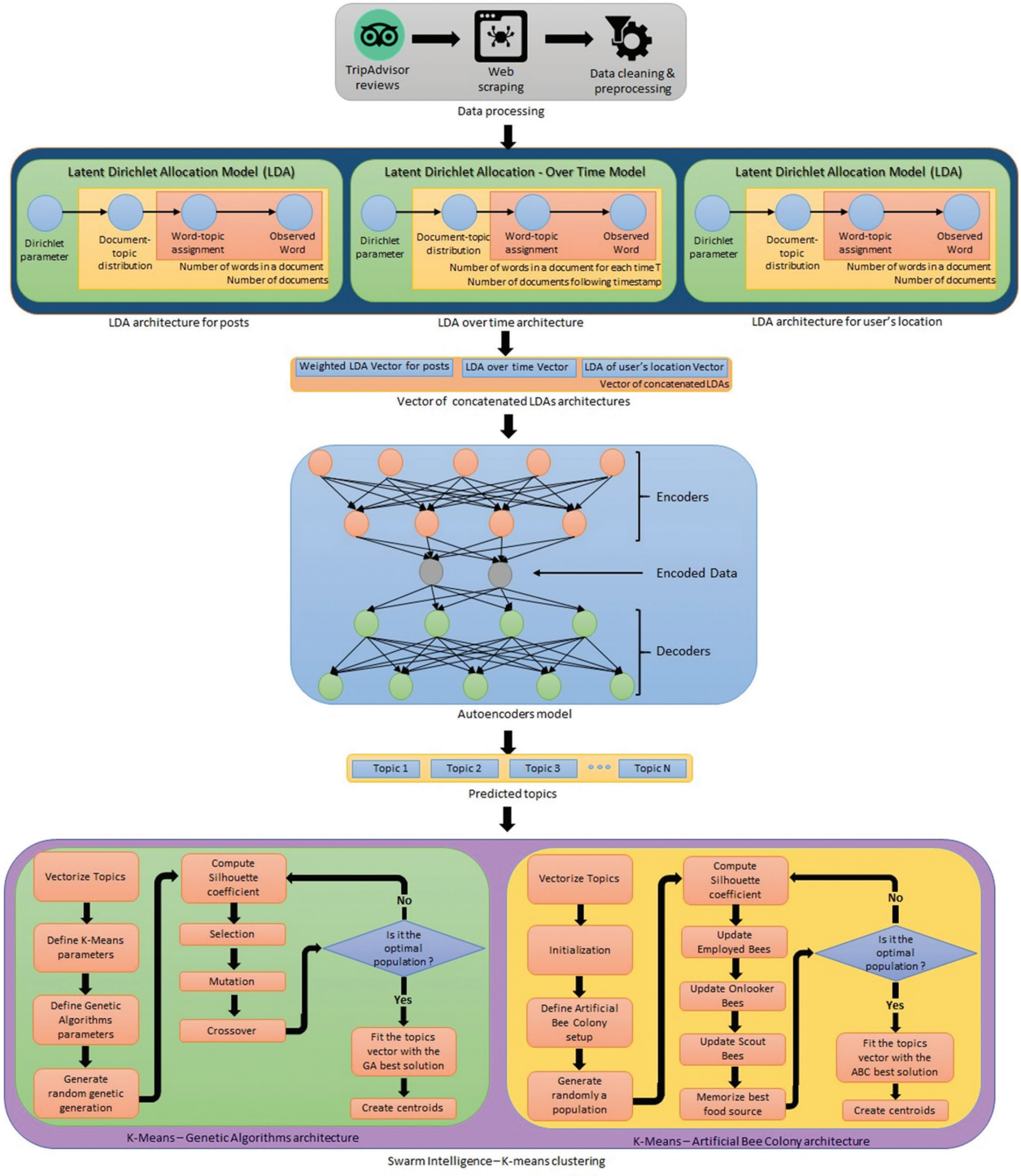

Fig. 1. MultLayered LDA approach with swarm intelligence clustering

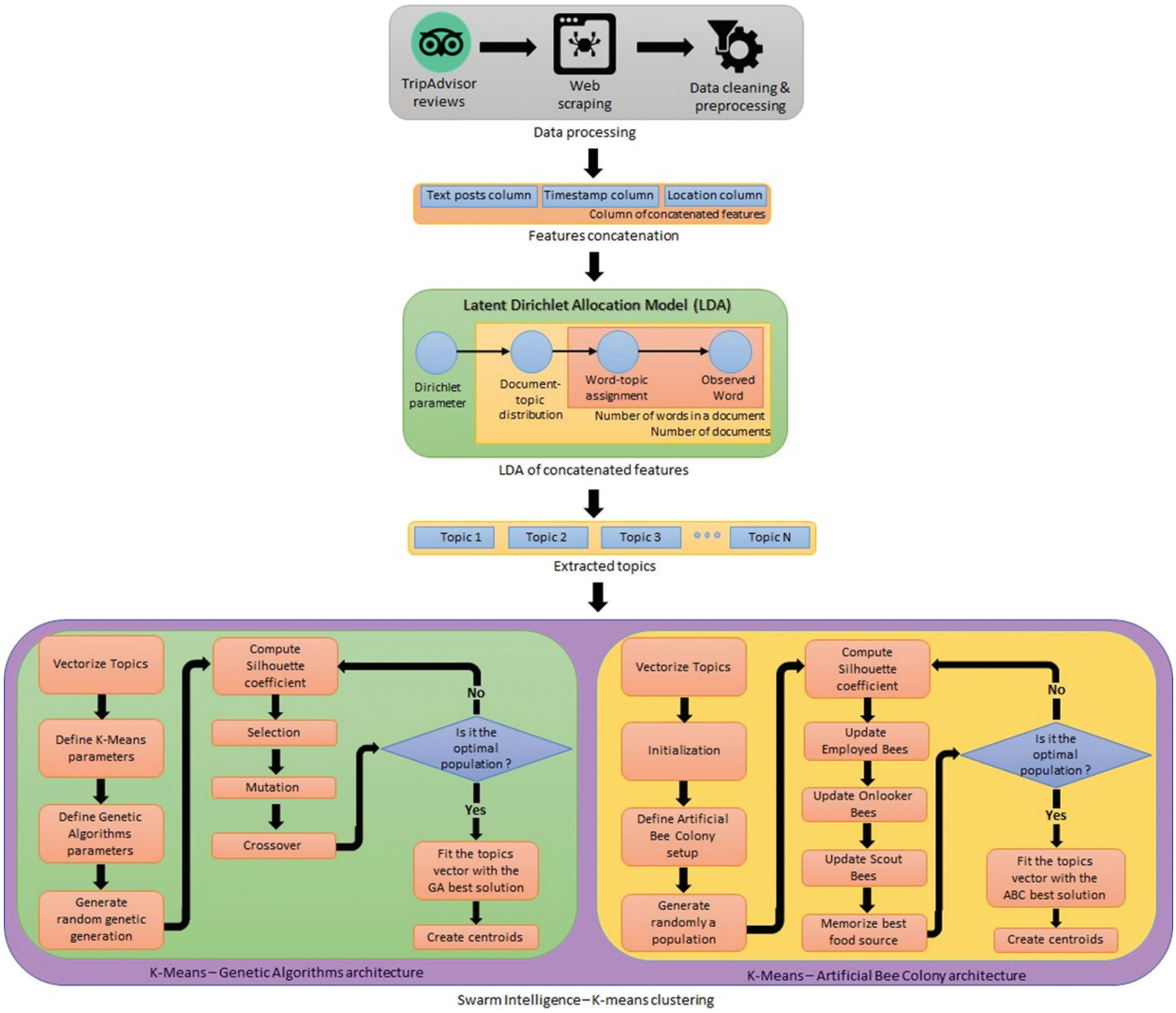

Fig. 2. Unified LDA approach with swarm intelligence clustering

4. RESULTS AND DISCUSSION

4.1. RESULTS

In this study, we conducted a comprehensive comparison of two different methods applied for Spatio-Temporal Topic Modeling, namely, Latent Dirichlet Allocation (LDA) and a hybrid approach that combines K-Means with Swarm Intelligence algorithms. These methods were applied to data derived from TripAdvisor, divided into three vectors: posts, the time of posts, and the location of the post's author. The performance of both methods was evaluated through several metrics, leading to compelling findings. The results from the two schemas provide interesting insights into the performance of different approaches to Spatio-Temporal Topic Modeling. Various metrics, including Coherence [18], Perplexity [19], and Topic Diversity [20] are used respectively for evaluating topic quality, and Silhouette Score [21], Davies-Bouldin Score [22], and Calinkski-Harabasz Score [23] to evaluate clustering performance. The various scores on the tables serve as metrics for the quality and performance of the applied Spatio-Temporal Topic Modeling approach. For all K-Means setups (i.e. Traditional K-Means and Swarm Intelligence K-Means), we set K to predict 20 clusters. The table below shows K-Means clustering with a loop over gamma values.

Table 1. K-Means clustering scores on Gamma values evaluation

Gamma Values	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
5	0.08	2.52	1445
10	0.23	1.54	4428
15	0.25	1.45	5023
20	0.26	1.20	4872
25	0.26	1.23	4834
30	0.25	1.46	4893
35	0.25	1.29	4915
40	0.26	1.45	4953
45	0.25	1.21	4949
50	0.25	1.46	4965

In Fig. 1., the Silhouette score, which provides a measure of how similar an object is to its own cluster

compared to other clusters, improves as the gamma value increases, with the highest score being 0.26 at a gamma value of 20. The Davies-Bouldin Score, indicative of intra-cluster similarity, decreases as the gamma value increases, suggesting that a higher gamma value results in more distinct clusters. Similarly, the Calinkski-Harabasz Score, indicative of the degree of separation between clusters, increased as the gamma value increased, suggesting more well-separated clusters with an increased gamma value.

Table 2. Multi-layered LDA topics quality

Model	Coherence	Perplexity	Topic Diversity
LDA for posts	0.53	-7.79	0.70
LDA over time	0.52	-7.64	0.68
LDA using author's location	0.51	-8.08	0.71

Examining the second table, we can see the performance of the LDA models for posts, LDA over time, and LDA using the author's location. The coherence score indicates that the topics generated were relevant and meaningful. This is particularly noticeable in the weighted LDA for posts, where the coherence score was the highest at 0.53. The Perplexity scores suggest a reasonable predictive performance of the models, with LDA using the author's location having the lowest Perplexity score of -8.08, indicating a better model. Lastly, the Topic Diversity values suggest a good spread of words across the identified topics, with LDA using the author's location demonstrating the highest Topic Diversity at 0.71. This suggests that this model was more successful in ensuring a broader range of topics.

Table 3. Multi-layered LDA clustering scores

	Models		
Metrics	K-Means	K-Means – Genetic Algorithms	K-Means – Artificial Bee Colony
Silhouette Score	0.26	0.26	0.35
Davies-Bouldin Index	1.20	1.24	0.81
Calinski-Harabasz index	4872	5047	44495

In examining the Silhouette Score, both the K-Means and Genetic Algorithms - K-Means methods achieved an identical value of 0.26, indicating a satisfactory degree of cluster compactness and separation, albeit not exceptional. In stark contrast, the application of the Artificial Bee Colony - K-Means approach demonstrated a substantial improvement, with a Silhouette Score of 0.35, reflecting superior cluster cohesion and separation. Further evaluation using the Davies-Bouldin Score showed a slight increase from 1.20 for K-Means to 1.24 when Genetic Algorithms were integrated, representing a moderate degree of cluster separation. However, the application of the Artificial Bee Colony - K-Means algorithm resulted in a score of 0.81, markedly lower than the aforementioned methods, signifying the provision of more distinct and well-separated clusters. With regard to the Calinski-Harabasz Score, the K-Means method yielded a score of 4872, experiencing a marginal improvement to 5047 when Genetic Algorithms were incorporated. Yet, when contrasted with these results, the Artificial Bee Colony - K-Means approach far surpassed both with a remarkable score of 44495. This substantial enhancement signifies that this method generates clusters that are not only denser but also more clearly separated.

These results suggest that using weighted LDA with optimized gamma values in Spatio-Temporal Topic Modeling can lead to significant improvements in topic relevance, predictability, and diversity in the tourism industry [24], which can be beneficial for gaining insights and trends from user-generated content. The following plots show the UMap projection of the predicted clusters using the swarm intelligence approaches:

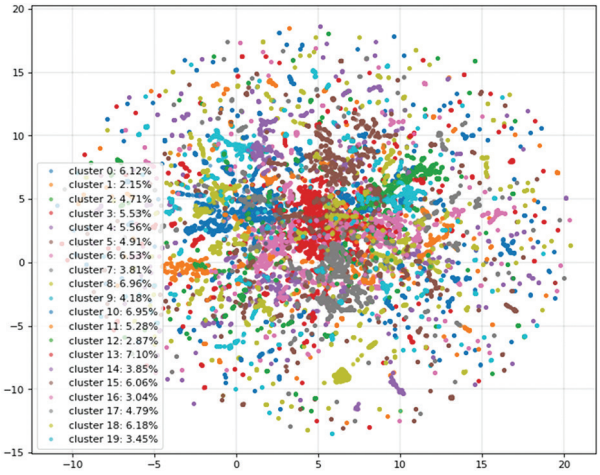

Fig. 3. MultiLayered LDA – Genetic Algorithms K-Means UMAP Projection

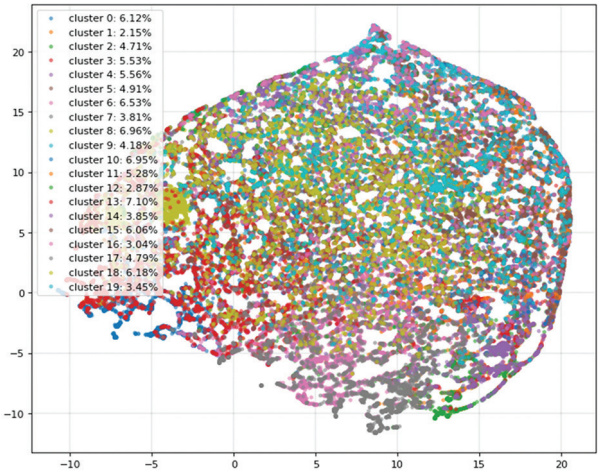

Fig. 4. MultiLayered LDA – Artificial Bee Colony K-Means UMAP Projection

The UMAP projection of the twenty predicted clusters offers illuminating insights into the relationships and dependencies within and across clusters. Particularly, the Artificial Bee Colony - K-Means (ABC-K-Means)

model (Fig. 4.) demonstrates a significant degree of interrelatedness and dependence among clusters. The proximity of some clusters and the absence of discernible boundaries in others suggest possible correlations and mutual influences between the topics within these groups. This pattern of clustering indicates that the ABC-K-Means model is adept at recognizing and incorporating the inherent relationships and dependencies in the data, thereby producing clusters that capture the multidimensional structure of the dataset. In contrast, the Genetic Algorithms - K-Means model (Fig. 3.) reveals a notably different pattern. The clusters in this projection appear to be more dispersed and independent, with clear demarcations separating the individual clusters. This spread signifies a higher degree of randomness in the distribution of topics across clusters [25]. The lack of apparent relationships or dependencies between clusters may suggest that this model tends to view each topic as an independent entity, leading to a more scattered and separated clustering [26]. It is therefore inferred that the Genetic Algorithms - K-Means model may be more suitable for datasets where topics are distinct and unrelated.

The following section presents an in-depth analysis and interpretation of the outcomes derived from the second method implemented in this study. The discussion that follows examines the effectiveness and the distinctiveness of this approach, drawing upon various evaluation metrics to assess the quality, coherence, and diversity of the generated topics

Table 4. Unified LDA topics quality

Model	Coherence	Perplexity	Topic Diversity
LDA for concatened features (Posts + Timestamps + Locations)	0.48	-7.49	0.49

In Method II, the LDA model's performance, in terms of Coherence and Perplexity scores, is slightly less optimal compared to Method I, with scores of 0.48 and -7.49, respectively. However, the Topic Diversity score is closely aligned with that of Method I, implying a robust distribution of words across the identified topics.

Table 5. Unified LDA clustering scores

	Models		
Metrics	K-Means	K-Means – Genetic Algorithms	K-Means – Artificial Bee Colony
Silhouette Score	0.25	0.78	0.72
Davies-Bouldin Index	1.27	0.6	0.61
Calinski-Harabasz index	5286	133744	122475

Traditional clustering methodologies like K-Means attain a marginally lower Silhouette Score of 0.25 compared to Method I. However, the incorporation of Genetic Algorithms and Artificial Bee Colony techniques substantially amplifies these scores, signaling the formation of better-defined and distinct clusters. Significantly, the Genetic Algorithms strategy achieved the highest Calinski-Harabasz Score, indicative of highly dense and well-separated clusters. To summarize, while both methods showcase impressive performances, their approaches to handling data diverge. Method I, deploying separate LDAs, presents slightly superior coherence and lower perplexity but lags in clustering metric performance. In contrast, Method II, despite demonstrating lower coherence and increased perplexity, stands out significantly in terms of its ability to form clusters, particularly when swarm intelligence algorithms are applied. The UMAP projections of the three clustering methodologies - K-Means, Genetic Algorithms-KMeans (GA-K-Means), and Artificial Bee Colony-K-Means (ABC-K-Means) - provide a visual insight into their effectiveness in data segregation. These projections, presented below, highlight the discernible differences between the approaches.

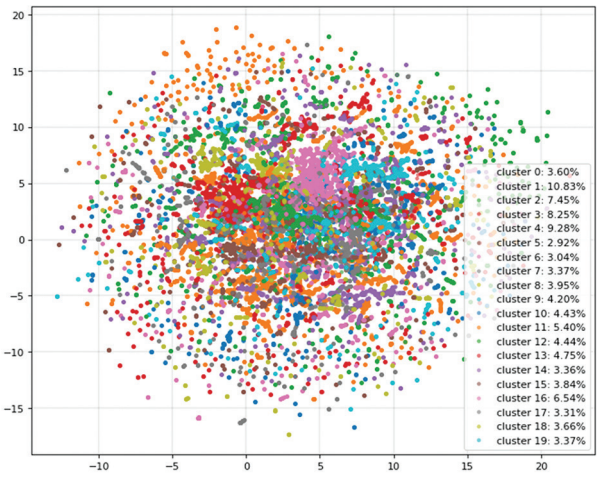

Fig. 5. Unified LDA –K-Means UMAP Projection

For the K-Means method (Fig. 5.), the clusters appear to be dispersed seemingly at random, with overlapping boundaries that suggest a lack of strong relationships between clusters.

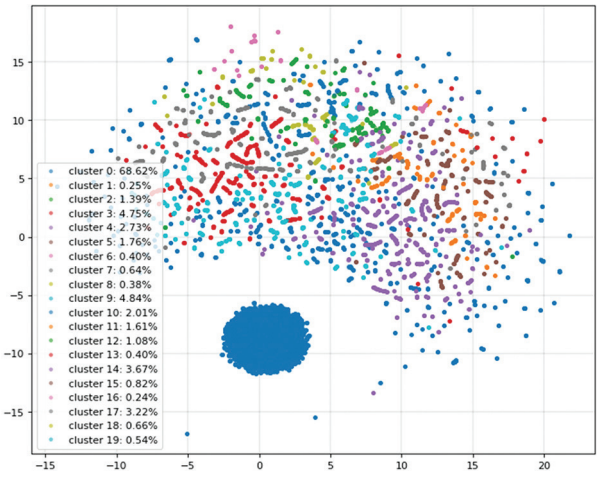

Fig. 6. Unified LDA – Genetic Algorithms K-Means UMAP Projection

In contrast, the UMAP projection for the GA-K-Means (Fig. 6.) method shows distinct clusters, suggesting a more structured segregation of data.

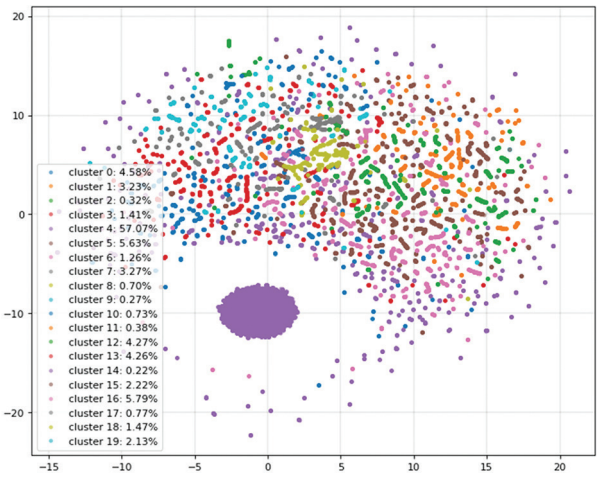

Fig. 7. Unified LDA – Artificial Bee Colony K-Means UMAP Projection

The ABC-K-Means method's projection (Fig. 7.) reveals a striking feature: a substantial cluster comprising a large number of samples, clearly differentiated from the remaining clusters. This dominant cluster's presence suggests potential relationships or dependencies between this prominent cluster and the other smaller clusters.

In the dynamic domain of the tourism industry, understanding the patterns and themes in tourists' discussions can provide crucial insights for decision-makers [27]. Word Cloud visualizations offer a potent tool to discern these patterns by prominently displaying the dominant terms within a text corpus. When applied to the clusters identified in our study, Word Clouds [28] can reveal the principal themes of each cluster, allowing us to ascertain the primary topics of discussion, temporal trends, and geographic references.

Considering the volume of data and the constraints of space in a research paper, presenting all 20 clusters for each methodology might not be feasible. Instead, a representative selection that conveys the distinctness and thematic richness of these clusters can effectively serve the same purpose.

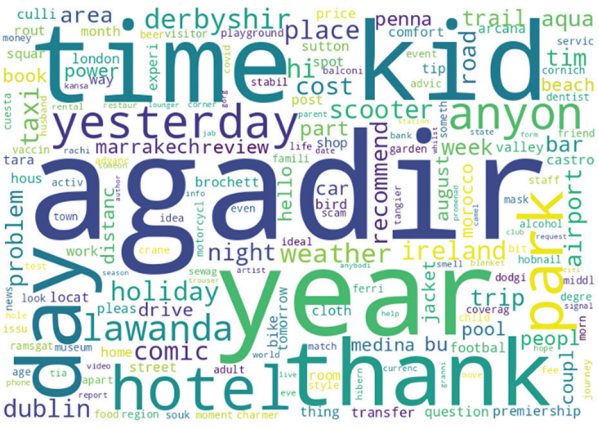

Fig. 8. WordCloud of Unified LDA-Swarm Intelligence (Cluster 4, Cluster 0)

For instance, consider the two representative clusters visualized above (Fig. 8.). They provide a snapshot of how each cluster encapsulates a theme interlacing location, time, and related discussions. Despite being just a fraction of the total clusters, these examples sufficiently showcase how each cluster signifies a unique theme, thereby illustrating the capabilities of this spatiotemporal topic modeling approach.

4.2. DISCUSSION

Our findings make a substantial contribution to the burgeoning field of Spatio-Temporal Topic Modeling, particularly demonstrating the strengths and potentialities of our two proposed schemas in efficiently extracting pertinent information from an extensive corpus of user-generated content. Each schema boasts unique attributes and yields notable advancements within this domain. When evaluated through the lens of our findings, Schema I, which independently applies LDA to the text, time, and location, produces comparatively higher Coherence scores. This indicates that the topics discerned through this method are not only more pertinent but also imbued with deeper meaning. This outcome resonates with prior research that accentuates the pivotal role of context in comprehending user-generated content [29]. Nonetheless, the enhanced clustering scores that emerge when Genetic Algorithms and Artificial Bee Colony are integrated underscore the value of utilizing swarm intelligence algorithms to navigate large data clusters. In contrast, (Fig. 2.), despite a slightly lower coherence and higher perplexity, delivers outstanding results in forming distinct and homogenous clusters. The algorithm's proficiency is particularly noticeable when swarm intelligence methodologies are incorporated. This indicates that a comprehensive overview of the generated topics can be obtained when context, in terms of location and time, is amalgamated with text into a single vector, leading to better-separated and more homogeneous clusters [30]. These findings have practical implications that stretch beyond the sphere of academic interest, specifically, within the tourism sector, ascertaining pat-

terns and trends in tourist behaviors and preferences is pivotal for strategic decision-making. Our models are engineered to extrapolate nuanced insights from the enormous and continually expanding corpus of user-generated content available on platforms such as TripAdvisor. In the context of Morocco's tourism sector, the Schema II model demonstrated substantial relevance. The word maps generated from this model gave a comprehensive overview of the discussion trends, incorporating time and location variables. The visually displayed results illustrate the relational dynamics within the tourism sector. The insights drawn from these maps can be leveraged by Moroccan tourism authorities to understand the temporal and spatial patterns of discussions and to strategize their services accordingly. The results of our research contribute to the expanding body of literature exploring advanced methodologies for topic modeling and clustering. The study underscores the significance of innovative methods in analyzing user-generated content and the potential of these approaches in extracting actionable insights across various sectors, including tourism.

5. CONCLUSION

In conclusion, our study introduces two novel schemas for Spatio-Temporal Topic Modeling using data from TripAdvisor's Morocco forum. We found distinct advantages in each schema, providing key insights into topic modeling. Schema I, which applies Latent Dirichlet Allocation (LDA) independently to text, time, and location, offered more coherent and predictable topics. Schema II, which integrates context and content into one vector, excelled in forming distinct clusters, especially with swarm intelligence algorithms like Genetic Algorithms.

These findings bear significant implications for tourism, enabling decision-makers to identify trends, enhance services, and create strategic plans based on tourist behavior. However, we acknowledge our study's limitations, including a reliance on TripAdvisor data and a fixed temporal scope (2007-2023). Future research will seek to integrate more data sources, apply advanced topic modeling techniques, and expand the temporal range to validate our schemas further.

6. REFERENCES

[1] S. Easton, N. Wise, "Online portrayals of volunteer tourism in Nepal: Exploring the communicated disparities between promotional and user-generated content", Worldwide Hospitality and Tourism Themes, Vol. 7, No. 2, 2015, pp. 141-158.

[2] Y. Liu et al. "Spatio-temporal topic models for check-in data", Proceedings of the IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14-17 November 2015, pp. 889-894.

[3] W. Min et al. "Multimodal spatio-temporal theme modeling for landmark analysis", IEEE MultiMedia, Vol. 21, No. 2, 2014, pp. 20-29.

[4] X. Luna et al. "Predictive spatio-temporal models for spatially sparse environmental data", Statistica Sinica, Vol. 15, 2005, pp. 547-568.

[5] J. Chen et al. "Local topic detection using word embedding from spatio-temporal social media", Proceedings of the International Conference on Neural Information Processing, 2019, pp. 629-641.

[6] K. Zhao et al. "Topic exploration in spatio-temporal document collections", Proceedings of the International Conference on Management of Data, San Francisco, CA, USA, June 2016.

[7] M. Asghari, D. Sierra-Sosa, A. S. Elmaghraby, "A topic modeling framework for spatio-temporal information management", Information Processing & Management, Vol. 57, No. 6, 2020, p. 102340.

[8] Y. Zhang, C. F. Eick, "Tracking events in Twitter by combining an LDA-based approach and a density-contour clustering approach", International Journal of Semantic Computing, Vol. 13, No. 1, 2019, pp. 87-110.

[9] C. González-Santos, M. A. Vega-Rodríguez, C. J. Pérez, "Addressing topic modeling with a multiobjective optimization approach based on swarm intelligence", Knowledge-Based Systems, Vol. 225, 2021, p. 107113.

[10] Q. Zhang, H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition", IEEE Transactions on Evolutionary Computation, Vol. 11, No. 6, 2007, pp. 712-731.

[11] E. Shabaninia et al. "Transformers in action recognition: A review on temporal modeling", abs/2302.01921, 2022.

[12] J. Diaz et al. "Spatio-temporal conditioned language models", Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 25-30 July 2020.

[13] D. Maier et al. "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology", Communication Methods and Measures, Vol. 12, No. 2-3, 2018, pp. 93-118.

[14] W. Wang et al. "Topic-guided variational auto-encoders for text generation", arXiv:1903.07137, 2019.

[15] S. Forrest, "Genetic algorithms", ACM Computing Surveys, Vol. 28, No. 1, 1996, pp. 77-80.

[16] D. Karaboga, B. Akay, "A comparative study of artificial bee colony algorithm", Applied Mathematics and Computation, Vol. 214, No. 1, 2009, pp. 108-132.

[17] Q. Gu, Z. Li, J. Han, "Linear discriminant dimensionality reduction", Proceedings of Machine Learning and Knowledge Discovery in Databases: European Conference, Athens, Greece, 5-9 September 2011, pp. 549-564.

[18] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Buttler, "Exploring topic coherence over many models and many topics", Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, July 2012, pp. 952-961.

[19] W. Zhao et al. "A heuristic approach to determine an appropriate number of topics in topic modeling", BMC Bioinformatics, Vol. 16, Suppl. 13, 2015.

[20] F. Nan, R. Ding, R. Nallapati, B. Xiang, "Topic modeling with Wasserstein autoencoders", arXiv:1907.12374, 2019.

[21] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, Vol. 20, 1987, pp. 53-65.

[22] D. L. Davies, D. W. Bouldin, "A cluster separation measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 2, 1979, pp. 224-227.

[23] T. Calinski, J. Harabasz, "A dendrite method for cluster analysis", Communications in Statistics - Simulation and Computation, Vol. 3, No. 1, 1974, pp. 1-27.

[24] Q. Li, S. Li, S. Zhang, J. Hu, J. Hu, "A review of text corpus-based tourism big data mining", Applied Sciences, Vol. 9, No. 16, 2019, p. 3300.

[25] W. Li, Y. Feng, D. Li, Z. Yu, "Micro-blog topic detection method based on BTM topic model and K-means clustering algorithm", Automatic Control and Computer Sciences, Vol. 50, 2016, pp. 271-277.

[26] M. S. Handcock, A. E. Raftery, J. M. Tantrum, "Model-based clustering for social networks", Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 170, No. 2, 2007, pp. 301-354.

[27] Z. Doborjeh, N. Hemmington, M. Doborjeh, N. Kasabov, "Artificial intelligence: a systematic review of methods and applications in hospitality and tourism", International Journal of Contemporary Hospitality Management, Vol. 34, No. 3, 2022, pp. 1154-1176.

[28] F. Heimerl, S. Lohmann, S. Lange, T. Ertl, "Word Cloud Explorer: Text analytics based on word clouds", Proceedings of the 47th Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 6-9 January 2014, pp. 1833-1842.

[29] L. Shifman, "Online entertainment | Cross-cultural comparisons of user-generated content: An analytical framework", International Journal of Communication, Vol. 10, 2016, p. 20.

[30] A. El-Kishky, Y. Song, C. Wang, C. Voss, J. Han, "Scalable topical phrase mining from text corpora", arXiv:1406.6312, 2014.