EAI Endorsed Transactions
on Intelligent Systems and Machine Learning Applications Research Article EAI.EU

BERTopic-Based Topic Modeling and Thematic Discovery in Long-Form Narrative Text

I.B.N HimaBindu [1,*], Sarojamma B.[2] and HaraGopal V.V.[3]

[1]Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India
[2]Department of Statistics, Sri Venkateswara University, Tirupati, Andhra Pradesh, India
[3]Department of Statistics, Osmania University, Hyderabad-7, Telangana, India

Abstract

With the increasing amount of digital text data available today, the demand for Natural Language Processing techniques is growing significantly. Topic modeling is a NLP technique for automatically identifying topics existing in a large corpus of text and deriving hidden patterns represented by that document collection, hence facilitating improved decision-making. The purpose of the present work is to explore the major topics of the renowned book “Autobiography of a Yogi”, written by Paramahansa Yogananda, an eloquent orator and a profound spiritual master. To accomplish the study, the most popular neural topic model ‘BERTopic’ was employed on the book. As a result, a number of intriguing topics are extracted, that are especially useful for those researchers and scholars delving into the complexities of the book as well as those interested in spirituality, Indian philosophy, the life journey and teachings of Paramahansa Yogananda.

Keywords: Textual data analytics, BERTopic, Topic modeling, topic probabilities, dimensionality reduction, HDBSCAN, UMAP

Received on 29 April 2026, accepted on 05 May 2026, published on 04 June 2026

doi: 10.4108/eetismla.12836

*Corresponding author. Email: himabindu.inampudi@gmail.com

1. Introduction

The major impetus for this work began with the question, "What is the grandeur of the world famous book “Autobiography of a Yogi” and why do those who have read it undoubtedly recommend it to others?”. This question was addressed in this paper by applying the well-known process of Topic modeling on the book “Autobiography of a Yogi” to extract the major topics from it without having to read it in its entirety.

1.1. “Autobiography of a Yogi”: A book by Paramhansa Yogananda

The book "Autobiography of a Yogi" is one of the most well-known autobiographies of its kind, and has set the spiritual path of many of its readers' lives. Paramhansa Yogananda, an eloquent orator and author, is the writer of the book "Autobiography of a Yogi". Swami Yogananda possessed a beautiful, narrative style that makes the book a pleasure to read. Numerous readers around the globe have adored this work. Even though it was first published in 1946, this book is still regarded as the best seller of all time. It continues to attract a growing number of readers. Over fifty languages throughout the world have a translation of this book. There are many examples of well-known people from all over the world who have read this book and as a result made profound spiritual changes in their lives.

1.2 Topic Modelling

Topic modelling is a statistical framework that assists users in comprehending huge document collections by discovering the abstract "topics" that are inherent in that corpus [1]. In Natural Language Processing, each word in the corpus is treated as a feature. Topic Modelling performs feature reduction and extracts only the needed attributes from the corpus thus allows us to focus on the relevant material rather than wasting time sifting through all of the text. This allows us to better understand the text and identify the hidden patterns which can be used to make predictions. Additionally, Topic Modelling can be used to identify the gist of a text, which can help us in summarizing the text quickly. Fig. 1 illustrates the topic modelling procedure.

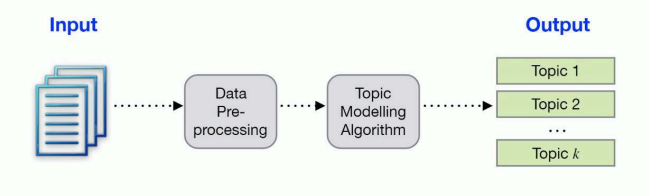

Figure 1. Topic Modelling process

1.3 BERTopic

"BERTopic is a topic modeling technique that leverages Bidirectional Encoder Representations from Transformers (BERT) embeddings and a class-based TF-IDF to create dense clusters" [2]. This technique has been used in this study to generate the topics from the text corpus.

2. Literature Review

In recent years, Topic Modelling has emerged as the method of preference for classifying enormous volumes of web material, Books, speeches, and discourses. The procedure of locating hidden themes inside a text's body and annotating those topics appropriately is referred to as "topic modelling." The literature has a number of topic modelling methodologies. As part of his study entitled "Probabilistic Latent Semantic Indexing", Hofmann [3] developed the EM algorithm in learning. The Latent Dirichlet Allocation (LDA) was proposed in 2003 by D. Blei et al [4] by revising the Probabilistic Latent Semantic Indexing model and learning framework via incorporating notions from the Bayesian approach there by fixing some of the previous method's limitations. While LDA uses a probabilistic approach, the Non-Negative Matrix factorization (NMF) technique uses matrix factorization approach [5]. The semantic interrelations between words are ignored in these approaches, as they describe each document as a bag- of - words. As a solution to this problem, text embedding methods have swiftly grown popular in the domain of natural language processing. With regard to producing contextual word- and sentence vector representations, "Bidirectional Encoder Representations from Transformers (BERT)" [6] and its variants (e.g.,[7]; [8]; [9]) have exhibited excellent performance. Although embedding approaches have been employed for several applications, from classification to neural search engines, topic modelling research has begun taking the benefits of these robust contextual representations. [10] showed that, as compared to traditional approaches like LDA, clustering embeddings with centroid-based algorithms are more effective at representing subjects. The efficacy of neural topic models in employing neural networks to advance current topic modelling methodologies has increased in recent years [11]; [12]; [13]; [14]. Recent years have seen an upsurge in topic modelling methods built entirely on embedding models. It serves as an example of the immense potential of embedding-based topic modelling techniques [15]; [16]; [17]. By clustering word and document embeddings, a number of techniques have begun to render the topic generation process simpler [10];[18]. Bertopic [2] is a technique based on clustering embeddings that uses a class-based variant of TF-IDF to construct topic representations.

3. Methodology

In BERTopic, topic representations are generated in three steps. In the first step, a pre-trained language model is used to convert each document into its embedding representation. In the second step, before the resulting embeddings are clustered, first they are reduced in dimensionality in order to optimize the clustering process. Finally, in third step, from these clusters of documents, topic representations are extracted. Workflow of the algorithm is shown in Fig. 2.

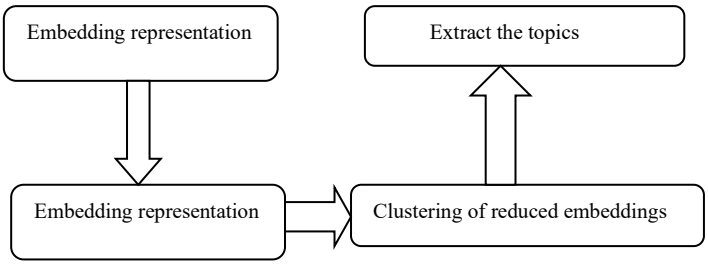

Figure 2. Workflow of the BERTopic algorithm

3.1 Document Embeddings

Semantic similarity is expected across documents covering the same topic in BERTopic. To compare the documents semantically, embedding the documents to create their representations in vector space takes place. This embedding

phase is carried out by BERTopic using the Sentence-BERT (SBERT) architecture [19];[20]. A dense vector representation of sentences, paragraphs and images can be obtained by using this framework that uses pre-trained language models. To achieve this task, in the present work, "all-MiniLM-L6-v2" language model that is trained specifically for semantic similarity tasks in English has been used.

3.2 Dimensionality Reduction

It has been shown that as the dimensionality of data increases, the distance to the nearest data point approaches the distance to the farthest data point [20];[21]. This leads to ill-defined spatial locality in high-dimensional space and distance measures are hardly different. As a means of combating this problem, dimensionality reduction techniques can be used to reduce the number of dimensions and make the embeddings more meaningful. BERTopic uses Uniform Manifold Approximation and Projection (UMAP), a popular non-linear dimensionality reduction method to minimize the dimensionality of embeddings produced in 3.1[22]. "There are three assumptions underlying this UMAP technique:

- There is a uniform distribution of data on a Riemannian manifold;
- The Riemannian metric is locally constant (or can be approximated as such);
- The manifold is locally connected.

Based on these assumptions, a fuzzy topological model can be developed for the manifold. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure."

3.3 Document Clustering

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is then used to cluster the reduced embeddings of 3.2.[23]. HDBSCAN extends DBSCAN by transforming it into a hierarchical clustering technique to detect clusters of varying densities. HDBSCAN employs a soft-clustering technique to construct clusters, enabling noise to be modelled as outliers. This keeps unrelated documents from being grouped together and is anticipated to enhance topic representations.

3.4 Topic Representations

The documents in each cluster are used to model the topic representations, with a topic being allocated to each cluster. On order to know what distinguishes one topic from another, based on the distribution of cluster words among each topic, a modified version of the popular metric TF-IDF can be adapted. The classic TF-IDF technique that reflects the significance of a word to a document can be modified thereby it may be used to describe the relevance of a term to a particular topic. "The conventional TF-IDF procedure combines two statistics, term frequency and inverse document frequency[24].

$$W_{t,d} = t f_{t,d} \log\left(\frac{\mathrm{N}}{\mathrm{df_t}}\right) \quad (1)$$

where term frequency (TF) is the frequency of term t in document d. The inverse document frequency (IDF), which is computed by taking the logarithm of the number of documents in a corpus N divided by the total number of documents that contain t, indicates how much information a word provides to a document.

This procedure is generalized here to clusters of documents. First, by simply merging the documents, we treat all documents in a cluster as a single document. Then, by converting documents to clusters, TF-IDF is modified to take into account this representation.

$$W_{t,c} = t f_{t,c} \log\left(1 + \frac{\mathrm{A}}{\mathrm{tf_t}}\right) \quad (2)$$

In this case, the term frequency in this situation reflects the frequent occurrence of term t in class C. In this case, the class C consists of clusters of documents that have been consolidated into a single document for each cluster. The inverse document frequency is then replaced by the inverse class frequency to determine how much knowledge a word adds to a class. It is derived by taking the logarithm of the average number of words per class A divided by the frequency of term t across all classes. We add one to the division within the logarithm to produce only positive values. In order to model the significance of terms in clusters rather than in individual documents, this class-based TF-IDF approach is very much useful. This enables us to produce topic-word distributions for each cluster of documents."

Finally, the total number of topics can be minimized to a user-mentioned number by continuously pairing the least prevalent topic's c-TF-IDF representations with its highly similar topic.

4. Results And Discussion

When implemented on our text corpus, the famous book" Autobiography of a yogi", Bertopic, by default, created hundreds of topics. Exploring so many different topics to get a fine-grained solution is a really challenging procedure. Therefore, based on the cosine similarity between c-TF-IDF vectors, we decreased the number of topics by combining pair of topics whose minimum similarity exceeded 0.9. After this step, from the extracted topics, some are randomly selected and presented here for better understanding.

4.1 Topic keywords

Some arbitrarily chosen topics among the extracted topics, as well as the top keywords corresponding to each are provided in Table I and Table II.

Table.1 Topics ---associated top keywords

Topic 20	Topic 24	Topic 11	Topic 19
Keywords	Keywords	Keywords	Keywords
'himalayas	'bose',	'ranchi'	'rabindranath tagore'
'kashmir',	'crescograph'	'high school'	'santiniketan'
'mountains'	'electron'	'selfrealization fellowship'	'poet'
'lakes',	'scientist',	'yogoda satsanga'	'bengali
'srinagar'	'microscope'	'kasimbazar'	'school',
'horse',	'plant'	'activities'	'literary'
'beauty',	'instrument'	'training'	'nobel',
'near',	'fern'	'taught'	'educational'
'mexico'	'cardiograph'	'educational'	'inspired',
'darker',	'tin'	'brahmacharya vidyalaya'	'received'
'steep',	'inventions'	'kasimbazar palace'	'gitanjali'
'tier'	'chloroform'	'lakshmanpur'	'received',
'peaks'	'physicist'	'maharaja'	'devendranath'
'reach himalayas'	'metals'	'ranchi students'	'lines',
'united provinces'	'new york'	'yogoda math'	'study'
'simla	'wireless'	'boys'	'song'
'grandeur'	'antidote'	'classes'	'singer'

From the resultant topics, many interesting insights can be extracted towards the book "**Autobiography of a Yogi"**. The gist of the resultant topics is presented below.

Swamy Yogananda was born to a Bengali household in Gorakhpur, India. His childhood name is MukundaLal Ghosh. His father worked for Bengal- Nagpur railway (Topic 8). His mother was an effectionate and charitable person by nature. They are the disciples of Shri.Lahiri Mahasaya of Varanasi (Topic 1). During his childhood, when attacked with cholera, in spite of trying their best, doctors could do nothing. Being asked by his mother, he gazed at the picture of Shri. Lahiri Mahasaya, for blessings (Topic 4). Instantaneously he witnessed a bright light enveloping his body which miraculously healed him (Topic 6).

Table 2. Topics ---associated top keywords

Topic 33	Topic 16	Topic 3
Keywords	Keywords	Keywords
'selfrealization'	'gandhi'	'kriya'
'church'	'mahatma'	'yoga'
'washington dc'	'nonviolence'	'technique'
'religions'	'satyagraha'	'yogi'
'mount'	'world'	'breath'
'los angeles'	'political'	'life'
'illustration'	'life'	'science'
'swami premananda'	'kasturabai'	'method'
'san diego'	'prison'	'practice'
'american'	'truth',	'mind'
'headquarters'	'india',	'ancient'
'selfrealization fellowship'	'means',	'patanjali'
'leader'	'make impossible'	'yogic'
'ranchi',	'diet',	'life force'
'building'	'leader'	'lahiri mahasaya'
'established'	'law'	'spiritual'
'hollywood'	'public'	'thoughts'

On later Occasions, while in meditating state, he experienced spiritual visions and found himself interacting with yogis faraway in the beautiful Himalayan foothills (Topic 20). As he grew older, so did his aspirations for the spiritual quest of finding his guru and visiting the sacred foothills of the Himalayas. He made multiple attempts planning his trip to Himalayas with his cousins and childhood friends but the efforts remained unsuccessful as either his brother Ananta would stand as a roadblock of his adventures or the trip would get cancelled due to some unforeseen events (Topic 30).

During this period, he came across several yogis, blessed with unfathomable powers (Topic 5) like Swamy Pranabananda (saint with two bodies), Gandh Baba (the perfume saint) (topic 23), Soham Swamy(Tiger swamy) (Topic 7), Nagendra Nath Bhaduri (the levitating saint) ,who all could make their powers possible by their severe practice of Yoga, Pranayama and meditation. Not only yogis and saints, but he also met one of the greatest scientists of modern Indian history, Jagadish Chandra Bose, who was the inventor of the crescograph and made revolutionary discoveries as a plant physiologist (topic 24).

In search of his ideal guru, he left to Banaras with his childhood friend Jitendra Mazumdar and finally met his guru shri Yukteswar Giri in Serampore, who initiated him into kriya yoga (topic 21). Shri Yukteswar Giri was honoured by the people with the title 'Gyanavathar'. His aura radiated wisdom of self-realization. His speeches carried intellect and his voice soothing with pleasantness (Topic 2). After receiving his degree from Calcutta

university (topic 9), he became a monk at a young age with the new name Yogananda, giving up his identity as Mukunda Lal Ghosh.

With the generous help of Maharaja of KasimBazar, Swamy Yogananda, turned the palace of KasimBazar into the Yogada Satsang Brahmacharya Vidyalaya, a boy's school (Topic 11). After two years of establishment of his school, Yogananda met Viswakavi Rabindranath Tagore in his school Shanthinikethan to mutually share the perspectives about their educational ideals in their schools (Topic 19).

After receiving an invitation to serve as a delegate representing India at the International Congress of religious liberals in America, Swamy Yogananda left for America to introduce Kriya yoga to its people, spreading the spiritual wisdom of the east (topic 40). He addressed the congress with his astonishing speech and also delivered many public lectures in the following years in every part of the country (Topic 42). He also established the American headquarters of the self-realization fellowship in Los Angeles, California (topic 33). Luther Burbank, a renowned American horticulturalist had a great admiration for Swamy Yogananda. He had been initiated into kriya yoga by Yogananda (Topic 22).

On his return trip to India, he visited Therese Neumann in Bavaria, who was a stigmatist and member of the third order of Saint Francis (Topic 13). Appreciating his efforts in the west, Shri Yukteswar ji , his guru, honoured Swamy Yoganada with the title of 'Paramahansa' (Topic 18). The devastating news of the sudden demise of his guru Shri Yukteswar Giri broke out during Yogananda's visit to Kumbhamela in Allahabad, Utterpradesh (Topic 28).

Before returning to America, Swamy Yoganada visited various parts of India and met great yogis like shri. Ramana Maharshi, Ananda Moyi Maa (Topic 27), Giri Bala (the saint who never eats)(Topic 32) and also Mahatma Gandhi ji in his Sabarmati ashram (Topic 16). After returning to America, Paramahansa Yoganada dedicated the rest of his life to the upliftment of the socities in both east and west by promoting spiritual harmony through his teachings of Kriya Yoga (Topic 29).

Kriya Yoga has become widely known in modern India through the instrumentality of Lahiri Mahasaya. Millenniums ago, the ancient Indian sage Patanjali, mentioned Kriya Yoga as a body discipline, mental control, and meditating on the Cosmic Sound of 'Aum' (Topic 3). It is a simple, psychophysiological technique that decarbonizes and re-oxygenates the human blood. To revitalise the brain and spinal centres, the extra oxygen's atoms are converted into life current. The skilled yogi transforms his cells into pure energy by preventing the buildup of venous blood, which slows or stops tissue degeneration. (Topic 0).

As topic modeling is a very subjective field, often, users find it difficult to validate their models. An efficient way to address this issue is to visualize the extracted topics and assessing whether they make sense or not.

To gain more insights from the topics generated and to compare topic representations to each other, bar charts are created for the topics generated. The bar charts for the first twelve topics, with the top five terms for each topic, are shown in Figure. 3.

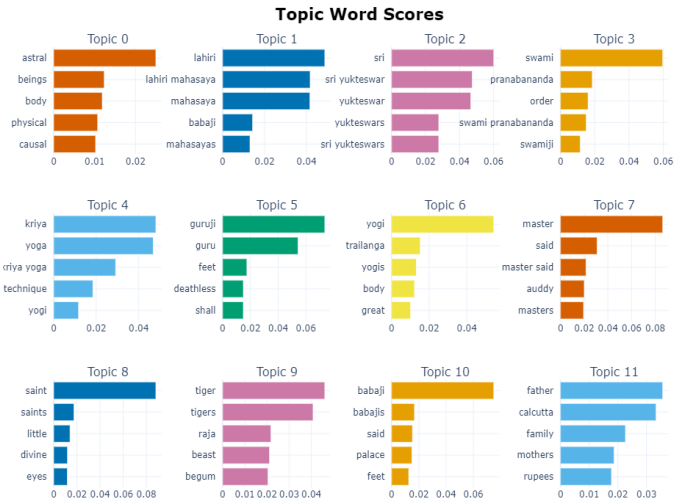

Figure 3. Bar charts of top keywords of some topics

We may iteratively go through the various topics after training our BERTopic model to acquire a thorough understanding of the extracted topics. But this process takes a long time and does not have a universal representation. Instead, we visualised the created topics by embedding our c-TF-IDF depiction of those topics in two-dimension using Umap and then visualized the two dimensions using Plotly, a Python open-source graphing tool. The resulting graphs are presented in Fig. 4 and Fig. 5**.**

The size of each circle in the above plot shows the frequency of the topic throughout all documents, while the circle itself represents a topic. If a topic is selected by using the slider, it turns red. Knowledge regarding a topic, such as its size and associated terms, is provided if we hover over it.

4.2 Topic probabilities

Top keywords from some randomly selected topics, together with their associated probabilities are shown in Table-3

The probability of a document pertaining to each conceivable topic is visualised for each document. Because there are numerous topics to visualise, the probability distribution of only the most probable topics is being visualised.

Figure 4. when topic 2 is selected

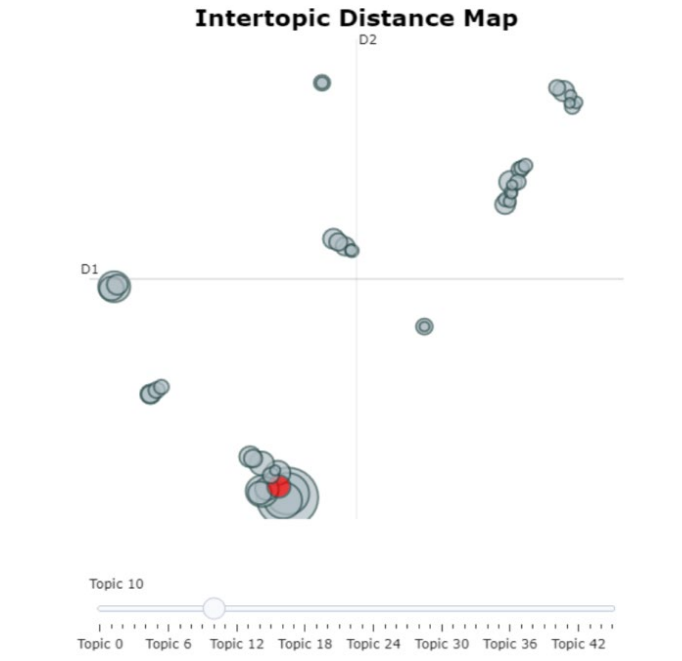

Figure 5. when topic 10 is selected

From figure-6 it can be interpreted that the model had some difficulties selecting the suitable topic for this document because there were multiple topics that were extremely similar to each other.

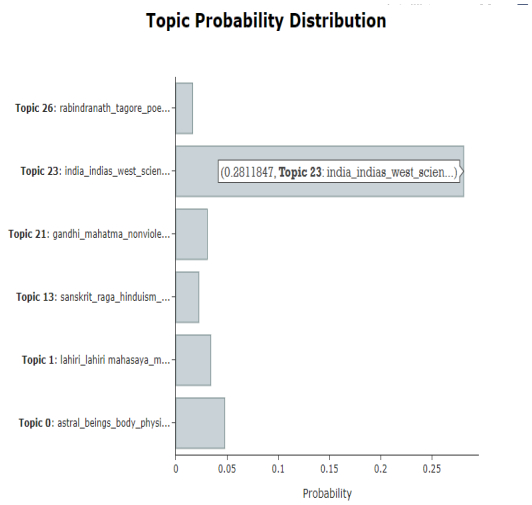

Figure 6. Topic probability distribution

To visualise the documents inside the topics and to determine whether or not they make sense, we used the plot presented in Fig.7 as a finer-grained method. For easy visualisation, this plot presents the document embeddings in a two-dimensional space.

The topics that were created can be hierarchically reduced. The plot presented in Fig. 8 is created in order to comprehend the probable hierarchical structure of the topics. It establishes the clusters and depicts their connections. When decreasing the number of topics generated, this enables us to choose a suitable number.

The findings from BERTopic might differ even if we execute the same code multiple times due to the stochastic nature of UMAP. Once the embeddings themselves are generated we experimented with BERTopic multiple times with varying parameters till we identified the topics that worked best for us. The 'topic coherence score' was then determined for number of topics created.

The level of semantic similarity between high-scoring terms in a specific topic is measured by Topic Coherence scores [1]. In topic modelling, this score is used to determine how interpretable the topics are to humans. We chose the well-known coherence metric referred to as the "C_v measure" from the several coherence metrics available. By leveraging their co-occurrences, it builds content vectors of words, and using Pointwise Mutual Information and cosine similarity, it determines the coherence score. For the number of topics generated, we achieved 0.3781 as the coherence score (Approximated to four decimal digits).

Based on the cosine similarity matrix between topic embeddings, a heatmap is created showing the similarity between topics and presented in Figure 8.

Table 3. Topic keywords ---associated Probabilities

Topic-33		Topic-3		Topic-24	
keyword	probability	keyword	Probability	keyword	Probability
'selfrealization'	0.0689535482028	'kriya'	0.0480397247478	'bose'	0.045717338529
'church'	0.0645003715981	'yoga'	0.0467974528165	'crescograph'	0.03845322015
'washington dc'	0.0274056257550	'technique'	0.0184271126314	'electron'	0.03369122667
'religions'	0.0520122866267	'yogi'	0.0116682980054	'scientist'	0.02403109300
'mount'	0.0315507176715	'breath'	0.0113275996821	'microscope'	0.02368992944
'los angeles'	0.0297213066432	'life'	0.0110957615867	'plant'	0.02088411201
'illustration'	0.0296796513327	'science'	0.0094181221972	'instrument'	0.02049392124
'swami premananda'	0.0285817775832	'method'	0.0089658097176	'fern'	0.01916783437
'san diego'	0.0264936148232	'practice'	0.0084851130909	'cardiograph'	0.019167834371
'american'	0.0246042044653	'mind'	0.0084216944184	'tin'	0.019167834371
'headquarters'	0.0245737885189	'ancient'	0.0070371289914	'inventions'	0.01776744708
'selfrealization fellowship'	0.0222909799828	'patanjali'	0.0069336445251	'chloroform'	0.01776744708
'leader'	0.0220098683499	'yogic'	0.0065205888855	'physicist'	0.017267869487
'ranchi',	0.0215952532858	'life force'	0.0064492395924	'metals'	0.016157514524
'building'	0.0215001238660	'lahiri mahasaya'	0.0064393063838	'new york'	0.013715201558
'established'	0.0212674467765	'spiritual'	0.0060169389797	'wireless'	0.013519921365
'hollywood'	0.0201599918864	'thoughts'	0.0057351240871	'antidote'	0.013519921365

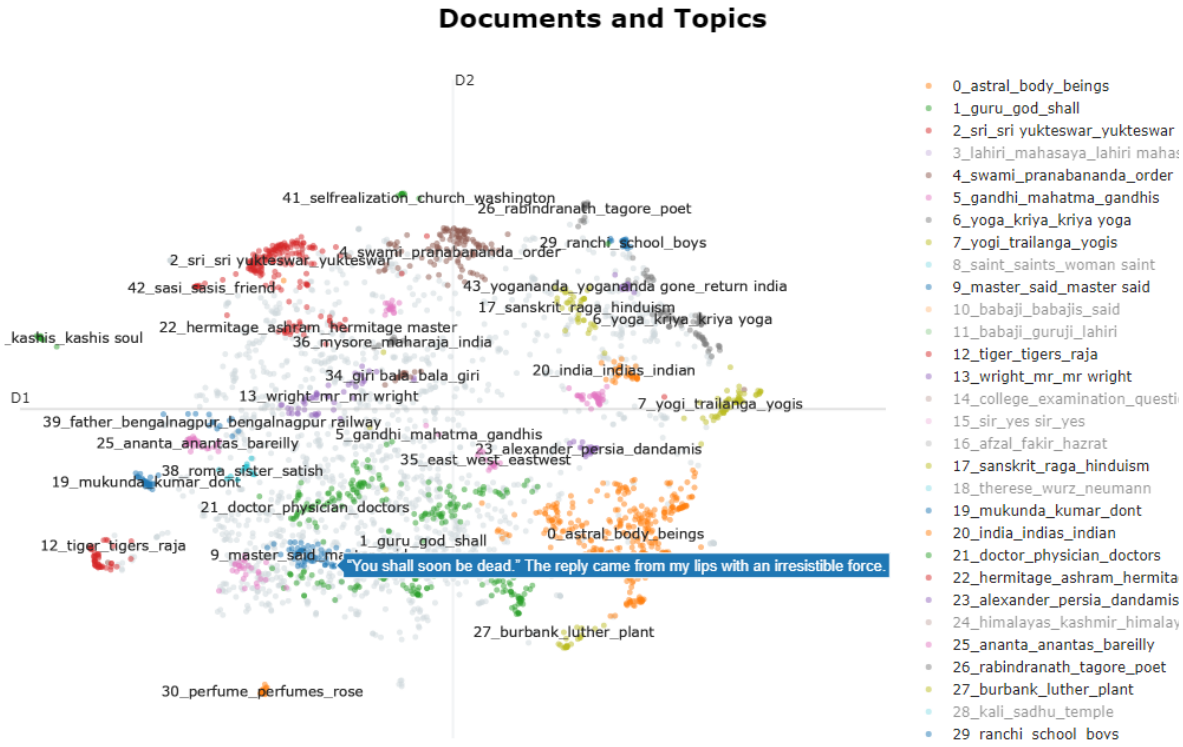

Figure 7. Visualization of documents inside the topics

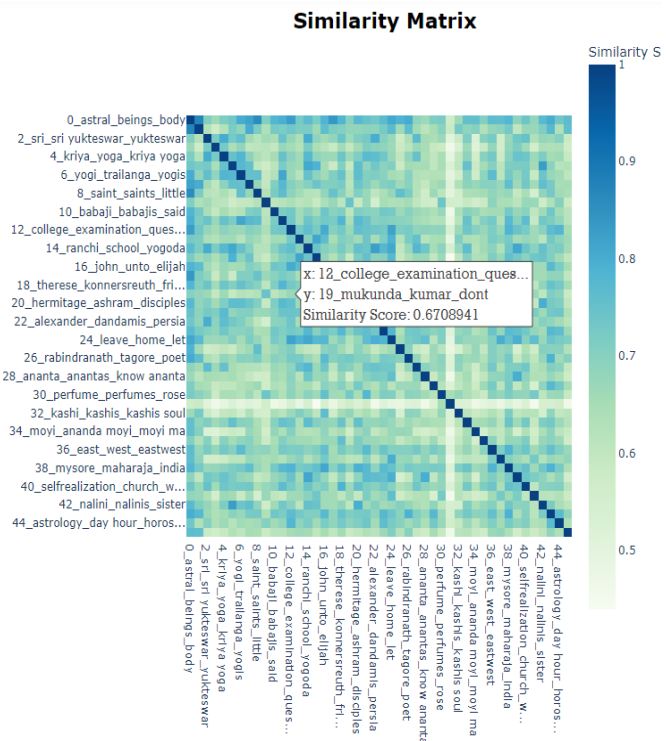

Figure 8. Heat Map

5. Conclusion And Future Work

In the present study, we employed topic modeling through BERTopic, a popular NLP technique to extract the essence of the renowned book "Autobiography of a Yogi". The topics extracted by the model includes the key themes of the power of meditation, the role of Guru-disciple relationship, the nature of consciousness, the yogic approach to health and healing and so on. According to the findings of this study, it can be concluded that the resultant topics presented a fairly appropriate overview of the book. We may also assert that application of topic modeling on literary works can help readers and researchers to better understand the content and ideas discussed within the book, and to gain insights into the author's perspective and worldview. As the future scope of this work, based on the specific usecases, one can customize the BERTopic model with embedding models like SpaCy, customized dimensionality reduction techniques like Principal Component Analysis and K–means algorithm for clustering rather than HDBSCAN.

References

[1] I. B. N. HimaBindu, S. V. Reddy, V. V. Haragopal, and B. Sarojamma, "Textual Analytics on 'Azadi Ka Amrit Mahotsav': Exploring Indian citizens' ideas for achieving Aatmanirbhar Bharat," in 2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Jan. 2023, pp. 1–8, doi: 10.1109/ICAECT57570.2023.10118308.

[2] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," Mar. 2022, doi: 10.48550/arXiv.2203.05794.

[3] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," pp. 177–196, 2001, [Online]. Available: https://doi.org/10.1023/A:1007617005950.

[4] D. M. Blei, A. Y. Ng, and M. T. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003, doi: 10.1162/jmlr.2003.3.4-5.993.

[5] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," Neural Computation, vol. 23, no. 9. pp. 2421–2456, 2011, doi: 10.1162/NECO_a_00168.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[7] J. Lee et al., "Data and text mining BioBERT : a pre-trained biomedical language representation model for biomedical text mining," vol. 36, no. September 2019, pp. 1234–1240, 2020, doi: 10.1093/bioinformatics/btz682.

[8] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," no. 1, Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.11692.

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS," in 8th International Conference on Learning Representations, ICLR 2020, 2020, pp. 1–17.

[10] S. Sia, A. Dalmia, and S. J. Mielke, "Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 1728–1736, doi: 10.18653/v1/2020.emnlp-main.135.

[11] S. Terragni, E. Fersini, B. Galuzzi, P. Tropeano, and A. Candelieri, "OCTIS: Comparing and optimizing topic models is simple!," EACL 2021 - 16th Conf. Eur. Chapter Assoc. Comput. Linguist. Proc. Syst. Demonstr., pp. 263–270, 2021, doi: 10.18653/v1/2021.eacl-demos.31.

[12] Z. Cao, S. Li, Y. Liu, W. Li, and H. Ji, "A novel neural topic model and its supervised extension," Proc. Natl. Conf. Artif. Intell., vol. 3, pp. 2210–2216, 2015, doi: 10.1609/aaai.v29i1.9499.

[13] H. Zhao, D. Phung, V. Huynh, Y. Jin, L. Du, and W. Buntine, "Topic Modelling Meets Deep Neural Networks: A Survey," in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, Aug. 2021, pp. 4713–4720, doi: 10.24963/ijcai.2021/638.

[14] H. L. and S. Lauly, "A neural autoregressive topic model.," Adv. Neural Inf. Process. Syst., 2012.

[15] F. Bianchi, S. Terragni, D. Hovy, D. Nozza, and E. Fersini, "Cross-lingual contextualized topic models with zero-shot learning," in EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 2021, pp. 1676–1683, doi: 10.18653/v1/2021.eacl-main.143.

[16] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," Trans. Assoc. Comput. Linguist., vol. 8, pp. 439–453, 2020, doi: 10.1162/tacl_a_00325.

[17] L. Thompson and D. Mimno, "Topic Modeling with Contextualized Word Representation Clusters," arXiv Prepr. arXiv2010.12626, Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.12626.

[18] D. Angelov, "Top2Vec: Distributed Representations of Topics," pp. 1–25, Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.09470.

[19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int.

Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 3982–3992, 2019, doi: 10.18653/v1/d19-1410.

[20] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, “On the surprising behavior of distance metrics in high dimensional space,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1973, no. February 2002, pp. 420–434, 2001, doi: 10.1007/3-540-44503-x_27.

[21] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is ‘nearest neighbor’ meaningful?,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 1540, pp. 217–235, 1998, doi: 10.1007/3-540-49257-7_15.

[22] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Feb. 2018, [Online]. Available: http://arxiv.org/abs/1802.03426.

[23] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” J. Open Source Softw., vol. 2, no. 11, p. 205, 2017, doi: 10.21105/joss.00205.

[24] T. Joachims, “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorixation,” Proc. ICML97, 1997.