Project Title

Automatic Topic Identification in Destination Review using Hybrid Topic Modelling Techniques

A PROJECT REPORT

Submitted by

BL.EN.U4CSE20092	M. Nitin Sai
BL.EN.U4CSE20114	N.Leeladhar Royal
BL.EN.U4CSE20113	N. Chakravarthi
BL.EN.U4CSE20110	N Siva Sai Raghu

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

AMRITA SCHOOL OF COMPUTING, BENGALURU

AMRITA VISHWA VIDYAPEETHAM

BENGALURU 560 035

NOVEMBER 2023

ACKNOWLEDGEMENTS

The satisfaction that accompanies successful completion of any task would be incomplete without mention of people who made it possible, and whose constant encouragement and guidance have been source of inspiration throughout the course of this project work.

We offer our sincere pranams at the lotus feet of **”AMMA”, MATA AMRITANANDAMAYI DEVI** who showered her blessing upon us throughout the course of this project work.

We owe our gratitude to **Prof. Manoj P.**, Director, Amrita Vishwa Vidyapeetham Bengaluru Campus. We thank **Dr. Sriram Devanathan,** Principal Amrita School of Computing, Bengaluru for his support and inspiration.

We would like place our heartfelt gratitude to **Dr. Gopalakrishnan E.A.,** Chair, Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru for his valuable support and inspiration.

It is a great pleasure to express our gratitude and indebtedness to our project guide: **Dr.Deepa Gupta, Professor,** Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru for her/his valuable guidance, encouragement, moral support, and affection throughout the project work.

We would like to thank express our gratitude to project panel members for their suggestions, encouragement, and moral support during the process of project work and all faculty members for their academic support. Finally, we are forever grateful to our parents, who have loved, supported and encouraged us in all our endeavors.

ABSTRACT

In a time when user-generated content is abundant on digital platforms, it is essential to analyze destination reviews in order to better understand tourist preferences and enhance the tourism experience.

This research proposes a novel method for automatically identifying and classifying topics within destination reviews using hybrid topic modeling techniques.

To extract latent topics from textual reviews, a methodology that incorporates Latent Dirichlet Allocation (LDA) and along with clustering is proposed. Moreover, the analysis is utilized to ascertain the degree of agreement or disagreement among viewpoints linked to every recognized subject, consequently augmenting the level of detail in the analysis.

The hybrid model's effectiveness is assessed with a large dataset of traveler reviews. The model's capacity to identify complex and contextually significant subjects inside reviews is demonstrated by the results, allowing for deeper insights.

This work advances the area by offering a strong foundation for automatically identifying topics in destination reviews, streamlining the decision-making procedures for tourism sector stakeholders, and promoting a more comprehensive comprehension of tourist preferences.

Keywords: Sentiment analysis, Latent Dirichlet Allocation, Topic modeling, Reviewing destinations, Tourism industry.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER - 1
INTRODUCTION

Understanding the opinions and tastes of tourists has become crucial in the era of digital platforms and growing user-generated content, thanks to the study of destination evaluations. The availability of a wide range of online evaluations and feedback makes it difficult to glean insightful information from this abundance of unstructured textual data. By using hybrid topic modeling techniques for the automatic identification and classification of subjects within destination reviews, this research offers a novel solution to this problem.

Travelers' varied experiences, viewpoints, and emotions are reflected in the wealth of information found in destination reviews. However, it takes a lot of effort and time to manually sort through this massive amount of data to find recurring patterns and subjects.

In order to tackle this issue, our research presents an advanced framework that combines the effectiveness of topic modeling approach—Latent Dirichlet Allocation (LDA) —to create a hybrid model specifically designed for destination review analysis. The justification for utilizing a hybrid method comes from the unique advantages and subtleties of various topic modeling approaches. By assuming a Dirichlet prior distribution over the document-topic and topic-word distributions, LDA, a probabilistic generative model, performs exceptionally well at detecting latent topics.

Our research attempts to overcome the shortcomings of individual strategies and utilize their synergies to obtain more accurate and nuanced subject identification within destination reviews by combining these methodologies into a cohesive hybrid model. Additionally, sentiment analysis is used to determine the degree of agreement or disagreement with each issue that has been found. This adds significant value to the investigation by illuminating visitors' attitudes and preferences with regard to particular elements of locations.

This study is important because it has the potential to transform the way destination reviews are analyzed, providing tourism industry stakeholders with an effective means of extracting valuable information from the large amount of unstructured textual data. By automatically recognizing and classifying themes, our framework aims to improve tourism experiences, expedite decision-making procedures, and promote a more in-depth understanding of travelers preferences.

The argument for using a hybrid method is based on the distinct advantages of several subject modelling approaches. Our study aims to overcome the limits of separate methodologies and capitalise on their synergies in order to obtain more accurate and nuanced subject identification within destination reviews.

This study's relevance goes beyond methodological innovation; it has the potential to revolutionise how destination reviews are properly analysed. Our approach aims to provide tourist industry stakeholders with an efficient and effective means of extracting important information from enormous troves of unstructured textual data through automated topic recognition and classification. The main goal is to not only improve tourism experiences and speed up decision-making processes, but also to foster a deep awareness of traveller preferences, so contributing to the worldwide refinement and advancement of the tourism industry.

CHAPTER - 2

LITERATURE REVIEW

1)The ACM Computing Surveys journal published a paper titled "Topic Modeling Using Latent Dirichlet Allocation: A Survey" in 2021. The authors, Umang Chauhan and Apoorva Shah, go into great detail about Latent Dirichlet Allocation (LDA) in the context of topic modeling. The importance of LDA in the field of Natural Language Processing (NLP) and its many uses in domains like text mining and information retrieval are covered in-depth in this survey.

Chauhan and Shah go into great detail to explain the basic ideas of LDA and go into detail about how it is used in document clustering and sentiment analysis. Though the paper provides a thorough explanation of the theoretical foundations and practical applications of LDA, it noticeably ignores the practical aspects that are crucial for practical implementation. Important real-world factors such as preprocessing methods and hyperparameter adjustments are absent from this paper.

2)A novel approach to literature recommendation is presented in the conference paper "A Novel Content-Based Recommendation Approach Based on LDA Topic Modeling for Literature Recommendation," written in 2021 by Bagul DV and Barve S. This method extracts topics from a literature dataset using Latent Dirichlet Allocation (LDA) topic modeling, then uses these topics to generate customized literature recommendations. The main goal is to improve the quality of recommendations by using the semantic content of documents in order to provide users with more individualized and useful recommendations that are in line with their preferences.

Although the paper offers a convincing approach for literature recommendation, there are significant research gaps that need to be addressed in the field. Notably, there is still a dearth of research in areas like scalability and evaluation metrics.

More research is necessary to address scalability issues with the method's effectiveness and performance when used on bigger or more varied datasets. Furthermore, the lack of strong assessment metrics prompts concerns regarding the efficacy and dependability of the technique in determining the caliber of recommendations.

This review of the literature emphasizes the exciting possibilities of content-based recommendation techniques, particularly when using LDA topic modeling to recommend literature. Nonetheless, it highlights the vital necessity of filling in research gaps, especially with regard to scalability issues and the creation of thorough assessment metrics, in order to guarantee the method's reliability and applicability in real-world scenarios.

3) The 2019 conference paper by Chen, B., Fan, L., and Fu, X. uses rules and the Latent Dirichlet Allocation (LDA) topic model to investigate sentiment classification in the tourism domain. This study looks into how these methods are used to analyze and classify sentiment in the tourism sector. Nonetheless, the review of the literature identifies gaps that demand attention. More advanced sentiment analysis techniques that are adapted to the unique subtleties of the travel industry are in demand. Furthermore, there hasn't been much research done on the practical difficulties that come with using sentiment analysis in the travel and tourism industry. These gaps highlight the need for additional research to improve sentiment analysis methods and deal with the particular difficulties of the tourism sector, thereby increasing the usefulness and effectiveness of sentiment classification.

4) Sethia K, Saxena M, Goyal M, and Yadav RK present a topic modeling framework that makes use of BERT, LDA, and k-means techniques in their conference paper from 2022. Through an exploration of the applications and synergies of these techniques, this framework seeks to advance subject modeling. However, the literature review reveals some significant gaps. The framework is devoid of comparative analyses with alternative methodologies and thorough evaluations of its efficacy. Furthermore, a more thorough understanding of its drawbacks and difficulties is required. Future research should concentrate on thorough exploration of challenges and constraints inherent in the integration of BERT, LDA, and k-means for topic modeling, as well as robust evaluations and comparative assessments against other methods, in order to strengthen the applicability of this framework. Closing these gaps would improve the framework's adaptability and dependability across a range of fields.

5) The limitations of mutual information (MI) and information gain (IG) in text classification are examined in Jian Yao and Jin Xu's paper from 2022. An example of a MI variation is IG, which concentrates on feature selection specifically, while MI measures the statistical dependence between variables. The literature points out the drawbacks of these text classification techniques. The survey highlights the need for more reliable and effective feature selection methods that are specific to tasks involving text classification. In addition, strategies for feature selection that take feature interaction into consideration are needed in order to get around the limitations and shortcomings of the current methodologies. Future research should concentrate on creating improved feature selection techniques that address these drawbacks and take feature interactions into account for more precise and effective text classification in order to advance text analysis systems.

6) Gao W, Fang Y, Li L, and Tao X's conference paper from 2021 uses graph neural networks (GNNs) to detect events in social media. This study uses graph neural networks (GNNs) to model social media data as a graph and identify relationships and patterns for event identification. There are obvious gaps in the paper, even though it probably describes the experiments, results, and methodology of using GNNs for event detection. The literature review emphasizes the need for thorough analyses that gauge the framework's efficacy, contrasts with other event detection techniques, and a more thorough understanding of the difficulties and constraints associated with applying GNNs in this field. Closing these gaps would improve the applicability and dependability of GNNs for event detection through in-depth assessments, comparative analyses, and a sophisticated comprehension of challenges in social media.

7) Computer Science Review published a paper in 2021 by Hamid, R. A., Albahri, A. S., Alwan, J. K., Al-Qaysi, Z., Albahri, O. S., Zaidan, A., Alnoor, A., Alamoodi, A. H., and Zaidan, B. that explores smart tourist recommendation systems, with a focus on e-tourism data management. It provides a thorough examination of data-driven tactics that improve the experiences of travelers and evaluates the most recent developments in smart tourism technology. But gaps in the literature are evident. Modern technology and the latest advancements in smart tourism must be included. Furthermore, a more thorough investigation of the difficulties associated with data management in e-tourism systems is imperative. Filling in these gaps would improve knowledge of modern smart tourism techniques and strengthen data management plans for improved experiences.

8) Specifically designed for unsupervised sentiment analysis tasks, the 2019 paper by Yifan Peng, Zhiyong Cheng, and Dan Yang, presented at the International Conference on Artificial Intelligence and Computer Science (ICAICS 2019), introduces novel improvements to the conventional Latent Dirichlet Allocation (LDA) model. The goal of this work is to improve topic modeling's understanding of sentiment by adding sentiment analysis capabilities to the traditional LDA framework. There are, however, some significant gaps in the literature. The study ignores consideration for more adaptable and open-ended approaches in favor of concentrating primarily on unsupervised learning paradigms. It emphasizes the necessity of efficient techniques for smoothly integrating sentiment analysis into LDA that are targeted toward unsupervised sentiment analysis in order to restrict investigation beyond this point. Closing these gaps would allow for a more thorough comprehension and use of sentiment-aware topic modeling techniques.

9) The GPT-2 (Generative Pre-trained Transformer 2) language model, functioning as an unsupervised multitask learner, is presented in the 2019 technical report by Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, published by OpenAI. Based on a transformer architecture, GPT-2 is trained on a wide range of language tasks without the need for task-specific supervision. It performs exceptionally well across a range of natural language processing benchmarks, demonstrating proficiency in tasks related to generation, translation, and language comprehension. Nevertheless, a thorough investigation of efficient fine-tuning techniques for particular tasks is absent from the paper. Moreover, the model's possible biases and scalability issues are not fully addressed. Optimizing these aspects would help unsupervised learning models like GPT-2 be as reliable and applicable as possible.

10) Presenting at the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT) in 2019, Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova introduced BERT, a groundbreaking language representation model that utilizes the transformer architecture. BERT is unique in that it takes into account both left and right contexts in every layer, allowing it to pre-train bidirectional representations from unlabeled text. It establishes the effectiveness of bidirectional context representations in language modeling with outstanding performance on a range of natural language understanding tasks.

The lack of a comparative analysis with other language representation models (LLMs) is a significant weakness in the paper, though. A thorough grasp of BERT's relative advantages and disadvantages in relation to other models is hampered by this lack of comparison, which is essential for determining its suitability and application in various applications involving natural language processing. Closing this disparity by means of comparative analyses would yield priceless information about the particular benefits and constraints of BERT relative to other LLMs.

CHAPTER – 3
SYSTEM SPECIFICATIONS

3.1 Software requirements

Natural Language Processing (NLP) Libraries:

Natural Language Processing (NLP) relies heavily on libraries, which provide crucial tools for language analysis and machine learning applications. NLTK (Natural Language Toolkit) distinguishes itself by providing a comprehensive suite of NLP tools, including tokenization, tagging, lemmatization, and part-of-speech tagging. With classes like LatentDirichletAllocation for topic modelling and KMeans for clustering, the sklearn package, which is best known for machine learning, contributes to NLP. Furthermore, the Gensim library specialises in LDA model development, using corpora to build the id2word Dictionary and a term corpus. These libraries, when combined, enable developers and researchers to effectively implement a wide range of NLP tasks, from fundamental linguistic analysis to complex machine learning applications such as topic modelling and clustering.

Deep Learning Libraries:

The Transformers library emerges as a critical tool in deep learning, specifically designed for natural language processing tasks. Pre-trained models such as GPT2Tokenizer and GPT2Model, which focus on transformer-based topologies, highlight its significance. Transformers, with a major focus on NLP, enables customers to seamlessly integrate sophisticated and cutting-edge language models into their applications or research projects.

Other Libraries:

Matplotlib is a core data visualization package that specializes in the development of plots and charts. Its versatility and ease of use make it an indispensable tool for visualizing data in a variety of forms. Other important libraries in the data science and machine learning sectors, in addition to Matplotlib, include torch, which is developed for tensor computation and deep learning applications, and NumPy, which is known for numerical calculation and efficient data processing. These libraries, when combined, form a strong toolkit for data scientists and researchers, allowing them to do tasks ranging from visualizing data trends to implementing complex machine learning models, hence improving the whole data analysis and modelling process.

CHAPTER - 4
SYSTEM DESIGN

High Level Design with description

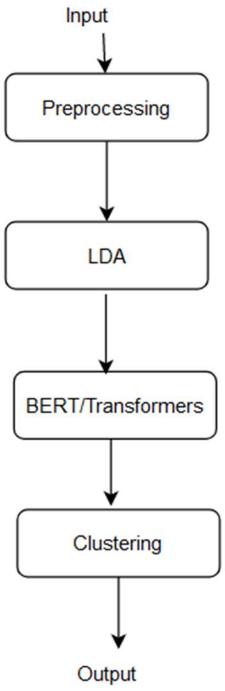


Fig 4.1: High level Design

The image's high-level design, as illustrated in Figure 4.1, can be summarized as follows: The input data is preprocessed to reduce noise and outliers and to convert it into a clustering-friendly format. To reduce the dimensionality of the data, Linear Discriminant Analysis (LDA) is performed. To extract features from text data, BERT/Transformers are used. The data is then grouped using k-means clustering .The procedure produces a set of clusters, each having a group of comparable data points.

Low Level Design with description

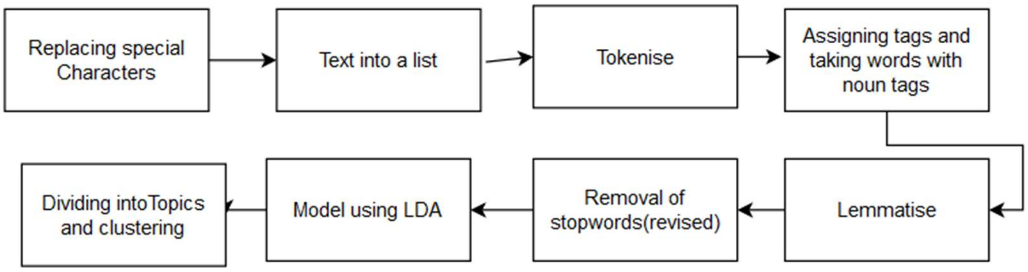

Replacing special Characters
Text into a list
Tokenise
Assigning tags and taking words with noun tags
Lemmatise
Removal of stopwords(revised)
Model using LDA
Dividing intoTopics and clustering

Fig 4.2: Low level design

The image's low-level design as illustrated in figure 4.2, depicts a multi-step method for improving a list of text by replacing special characters, tokenizing the text, giving tags to tokens, removing stop words, lemmatizing words, using topic modelling for list division, and clustering topics. Each step necessitates algorithmic decisions, such as the use of regular expressions or language-aware tokenizers. The selection of efficient algorithms is dictated by system performance and scalability requirements, particularly for processing huge volumes of data. Furthermore, the design must take into account the type of data, ensuring that algorithms are suited to text processing if necessary. The final result is a list of clustered topics extracted from the input text, with arrows representing data flow between steps in the flow diagram.

CHAPTER – 5
SYSTEM IMPLEMENTATION

5.1 Modules used with description:

1. LDA: It is a probabilistic generative model that depicts documents as combinations of topics. The underlying premise is that every document within a corpus is composed of a limited number of subjects, with each word inside the text being associated with a particular topic. Reverse-engineering this process, the model aims to deduce the underlying topics from the word distribution across papers.

2.Transformers (Hugging Face):

Description: Hugging Face created the open-source Transformers library, which offers pre-trained models and an easy-to-use interface for tasks related to natural language generation and understanding (NLU and NLG, respectively). Built on top of PyTorch and TensorFlow, it provides a vast array of pre-trained models for tasks like text classification, question answering, language translation, and text generation, including BERT, GPT.

3. Python LDAvis:

Description: A very helpful Python library for Latent Dirichlet Allocation (LDA) visualization is called pyLDAvis. It produces interactive visuals that facilitate investigating and understanding the subjects produced by LDA. It offers a means of dynamically investigating the subjects, connections between them, and word distribution within them.

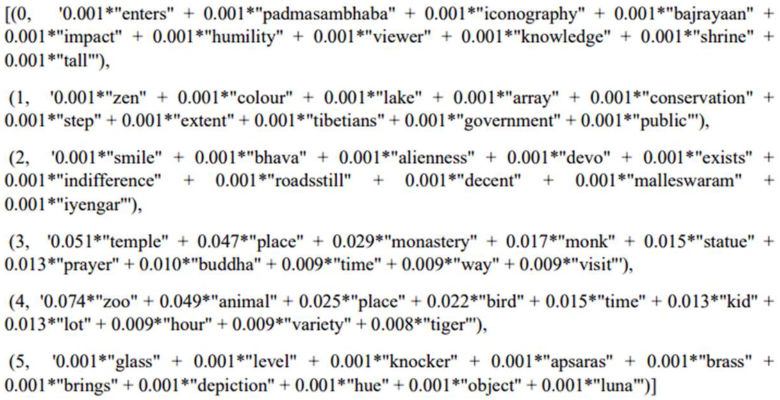
4.The Natural Language Toolkit (NLTK)

Description: A complete platform for developing Python applications that interact with data in human languages is called NLTK. It offers user-friendly libraries and interfaces for a variety of operations, including parsing, tagging, tokenization, and stemming. For problems involving natural language processing, NLTK offers a variety of corpora, lexical resources, and modules.

In the field of natural language processing, each of these libraries has a distinct function and set of tools that help researchers and developers create NLP applications, conduct analysis, and work effectively with text data.

CHAPTER - 6
SYSTEM TESTING

The Dataset taken for training is robust to outliers and hence training of LDA model provides the topic distribution, which is a good overview of the different types of locations and experiences that are reviewed by tourists in Mysore. The specific keywords associated with each topic can be used to gain deeper insights into the different aspects of these locations and experiences that are most important to tourists.



```
[(0,   '0.001*"enters" + 0.001*"padmasambhaba" + 0.001*"iconography" + 0.001*"bajrayaan" +
0.001*"impact" + 0.001*"humility" + 0.001*"viewer" + 0.001*"knowledge" + 0.001*"shrine" +
0.001*"tall"'),

(1,  '0.001*"zen" + 0.001*"colour" + 0.001*"lake" + 0.001*"array" + 0.001*"conservation" +
0.001*"step" + 0.001*"extent" + 0.001*"tibetians" + 0.001*"government" + 0.001*"public"'),

(2,  '0.001*"smile" + 0.001*"bhava" + 0.001*"alienness" + 0.001*"devo" + 0.001*"exists" +
0.001*"indifference"  +  0.001*"roadsstill"  +  0.001*"decent"  +  0.001*"malleswaram"  +
0.001*"iyengar"'),

(3,  '0.051*"temple" + 0.047*"place" + 0.029*"monastery" + 0.017*"monk" + 0.015*"statue" +
0.013*"prayer" + 0.010*"buddha" + 0.009*"time" + 0.009*"way" + 0.009*"visit"'),

(4, '0.074*"zoo" + 0.049*"animal" + 0.025*"place" + 0.022*"bird" + 0.015*"time" + 0.013*"kid" +
0.013*"lot" + 0.009*"hour" + 0.009*"variety" + 0.008*"tiger"'),

(5,  '0.001*"glass" + 0.001*"level" + 0.001*"knocker" + 0.001*"apsaras" + 0.001*"brass" +
0.001*"brings" + 0.001*"depiction" + 0.001*"hue" + 0.001*"object" + 0.001*"luna"')]
```

Fig 6.1: Topic Distribution after training LDA

Topic Distribution:

Instruction (bhava, indifference, smile, devo)

Zoo (animal, place, bird, time)

Hamlet (Zen, color, questioning, monk)

Temple (monastery, place, monk, statue)

Statue (devo, time, array, place)

Performance Evaluation Metric:

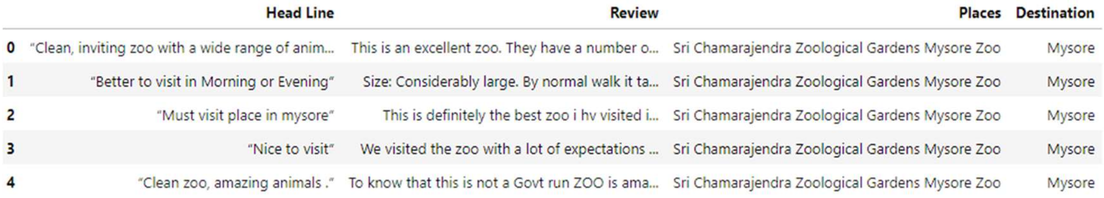
Coherence Score is used to assess the quality of extracted topics and make sure they are comprehensible and interpreted.

CHAPTER – 7

RESULTS AND ANALYSIS

Task 1: Cleaning Data

The dataset illustrated in fig 7.1, we carried out data cleaning, which included eliminating redundant reviews, dealing with missing numbers, fixing formatting problems, getting rid of unnecessary content, and normalizing casings. With regard to the locations "Sri Chamarajendra Zoological Gardens Mysore Zoo," "Avadhoota Datta Peetham," and "Namdroling Nyingmapa Monastery," the destination Mysore, the dataset currently has 1492 rows.



	Head Line	Review	Places	Destination
0	“Clean, inviting zoo with a wide range of anim...	This is an excellent zoo. They have a number o...	Sri Chamarajendra Zoological Gardens Mysore Zoo	Mysore
1	“Better to visit in Morning or Evening”	Size: Considerably large. By normal walk it ta...	Sri Chamarajendra Zoological Gardens Mysore Zoo	Mysore
2	“Must visit place in mysore”	This is definitely the best zoo i hv visited i...	Sri Chamarajendra Zoological Gardens Mysore Zoo	Mysore
3	“Nice to visit”	We visited the zoo with a lot of expectations ...	Sri Chamarajendra Zoological Gardens Mysore Zoo	Mysore
4	“Clean zoo, amazing animals .”	To know that this is not a Govt run ZOO is ama...	Sri Chamarajendra Zoological Gardens Mysore Zoo	Mysore

Fig 7.1: Dataset

Task 2: Extracting Review Text and Normalizing Casings

After standardizing the casings, we took the review text out of the dataset. To make it easier to manipulate in later jobs, the text was converted to lowercase, the review content was isolated from any formatting tags or information, and it was saved as a list of individual reviews.

Task 3: Tokenization

NLTK's word_tokenize function, which divides the review text into discrete words or tokens, to tokenize the reviews. This stage readies the text for subsequent processing operations including stop word removal, lemmatization, and stemming.

Task 4: Lemmatization and Stemming

By eliminating suffixes and prefixes, we reduced words to their basic forms using Porter's Stemming technique. This stage facilitates the grouping of relevant words together and enhances the performance of tasks that follow, such as topic modeling.

Task 5: Eliminating Stop Words

Common stop words were eliminated from the tokenized text. Stop words are words like articles, prepositions, and pronouns that have little to no relevance for topic modeling or sentiment analysis. Eliminating stop words improves the attention on more significant words while lowering the dimensionality of the data. The significant words that were derived are shown in fig 7.2.

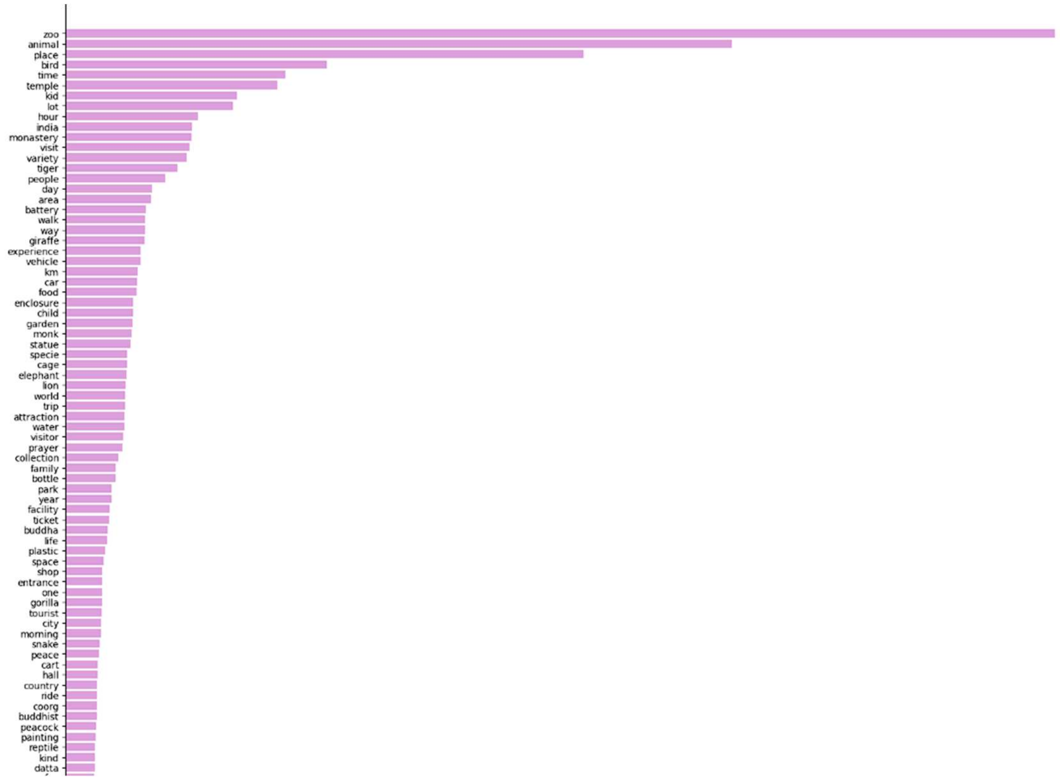


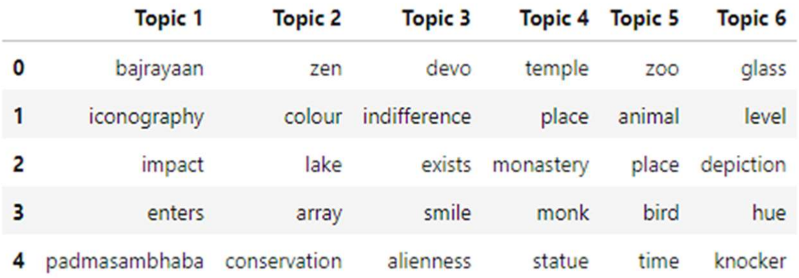
Fig 7.2: Bar plot to visualize the top common words

Task 6: Revising the Stop words List

We improved the list of stop words by taking into account the particular features of the dataset. This required looking at word frequencies and determining which words are commonly used but not semantically meaningful. To guarantee that the analysis concentrates on the most illuminating words, the stop words list should be refined.

Task 7: Topic Modeling Using LDA

Throughout the collection of reviews, we used Latent Dirichlet Allocation (LDA) to find latent themes or subjects. Using co-occurrence patterns as shown in fig 7.3, a statistical technique that classifies words into subjects. This stage aids in identifying the recurrent themes and emotions that are expressed in the reviews.



	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
0	bajrayaan	zen	devo	temple	zoo	glass
1	iconography	colour	indifference	place	animal	level
2	impact	lake	exists	monastery	place	depiction
3	enters	array	smile	monk	bird	hue
4	padmasambhaba	conservation	alienness	statue	time	knocker

Fig 7.3: Topic Distribution table

Task 8: Calculating the Coherence Score

To assess the caliber of the selected subjects, we computed coherence ratings. Coherence ratings assess a topic's interpretability and semantic consistency. More significant and well-defined subjects are indicated by higher coherence ratings.The initial experiments with the model yielded a coherence score of 0.54 for Six topics.

Task 9: Word Cloud Display

Word clouds are a valuable tool for analyzing LDA data. They can be used to determine which themes are the most crucial, comprehend the connections between various topics, and follow a topic's development over time. As illustrated in Fig7.4, Word cloud show the most frequently used terms related to the subject of "iconography,temple, place, monastery, monk, statue, prayer, buddha, time, way, visit". The most significant elements of this subject are represented by the largest words in the word cloud, and the connections between various terms shed light on how this issue has changed over time.

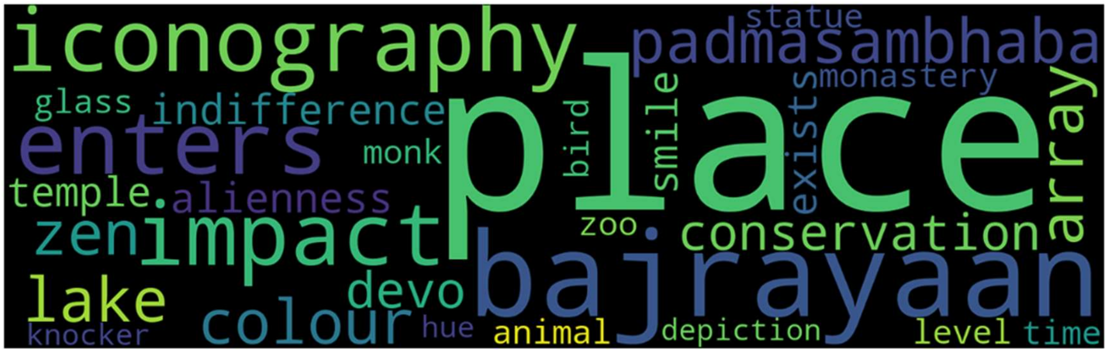


Fig 7.4: Word Cloud

Task 10: Clustering

The optimal number of clusters cannot be determined by the K-means method. Thus, the Elbow technique uses the K-means method to find the optimal number of clusters. The elbow technique can yield the same number of clusters K on the quantity of varied data, based on the outcomes of the process of calculating the optimal number of clusters. The default for the characteristic process will be the outcome of using the elbow approach to determine the optimal number of clusters. The elbow method graph is shown in the fig 7.5, we can interpret that the **optimal number of clusters to be selected are 3.**

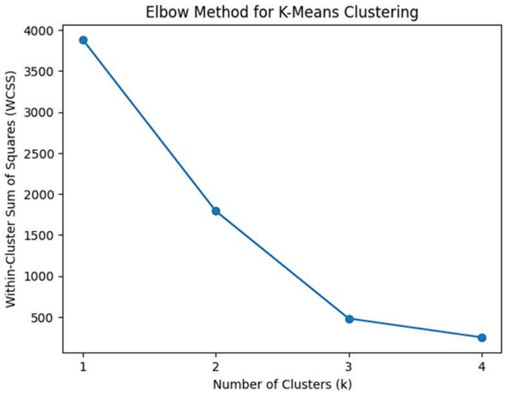


Fig 7.5: Elbow method for k-means Clustering

Task 11: LDA+BERT Embeddings or LDA+GPT-2

To improve the representation of the review text, we investigated using BERT or GPT-2 embeddings. These methods translate words into vector representations that describe word semantic associations. To further evaluate the topic modeling process, we introduced the silhouette score, which measures the separation between clusters. The silhouette score for K-means clusters was 0.08.

Clustering assignment results in the case of gpt-2 is shown in the fig 7.6 :

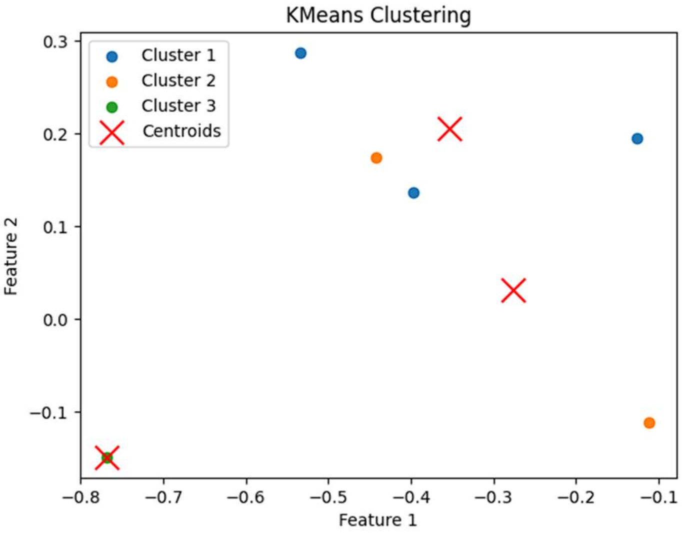


Fig 7.6: Visualizing k-means Clustering for LDA+BERT

We next replaced BERT embeddings with GPT-2 embeddings and a silhouette score of 0.46 for six topics. These scores are both higher than the scores obtained with BERT embeddings, indicating that GPT-2 embeddings can improve the quality of top ic modelling. These findings suggest that GPT-2 embeddings provide a promising app roach for enhancing the quality of topic modelling, potentially leading to more meani ngful and interpretable results.

Clustering assignment results in the case of gpt-2 is shown in the fig 7.7 :

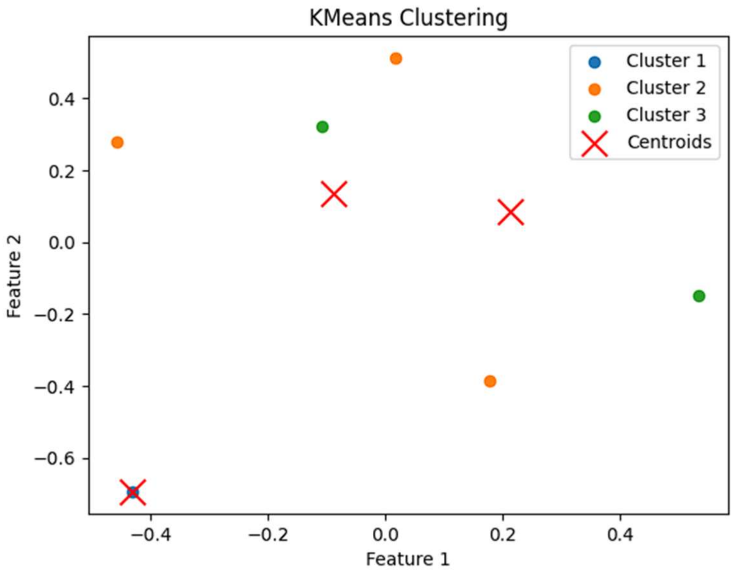


Fig 7.7: Visualizing k-means Clustering for LDA+GPT-2

CHAPTER – 8
CONCLUSION AND FUTURE SCOPE

To increase the precision and effectiveness of topic identification, it is necessary to investigate the efficacy of further sophisticated clustering algorithms or metaheuristic optimization strategies. Insights from the vast amount of traveler reviews available on social media platforms would be more trustworthy and instructive for researchers and decision-makers as a result. This would also help with more precise topic identification and clustering within the tourism domain, improving decision-making processes for travelers and travel operators alike. Reviews can be further evolved with the help of topic identification as base. With our model as base the reviews can be further divided into topics like cleanliness or taste of food from the general star rating which isn't as descriptive, further helping in the identification of related reviews of the user interested area.

REFERENCES

[1]Chauhan U, Shah A (2021) "Topic modeling using latent Dirichlet allocation: a survey." ACM Comput Surv (CSUR) 54(7):1–35.

[2]Hamid, R. A., Albahri, A. S., Alwan, J. K., Al-Qaysi, Z., Albahri, O. S., Zaidan, A., Alnoor, A., Alamoodi, A. H., & Zaidan, B. (2021). How smart is e-tourism? A systematic review of smart tourism recommendation system applying data management. Computer Sci Rev, 39, 100337.

[3] Chen, B., Fan, L., & Fu, X. (2019). Sentiment classification of tourism based on rules and LDA topic model. In 2019 International Conference on Electronic Engineering and Informatics (EEI) (pp. 471–475). IEEE.

[4] Yao J, Xu J (2022) "English text analysis system based on genetic algorithm." Mob Inf Syst. https://doi.org/10.1155/2022/9382890

[5] Sethia K, Saxena M, Goyal M, Yadav RK (2022) "Framework for topic modeling using BERT, LDA and k-means." In: 2022 2nd International conference on advance computing and innovative technologies in engineering (ICACITE), pp 2204–2208.

[6] Event Detection in Social Media via Graph Neural NetworkWeb Information Systems Engineering – WISE 2021, 2021, Volume 13080ISBN : 978-3-030-90887-4Wang Gao, Yuan Fang, Lin Li, Xiaohui Tao.

[7] Osmani, Amjad & Bagherzadeh, J. & Soleimanian Gharehchopogh, Farhad. (2020). Enriched Latent Dirichlet Allocation for Sentiment Analysis. Expert Systems. 37. 10.1111/exsy.12527.

[8] *Devlin, Jacob et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” North American Chapter of the Association for Computational Linguistics (2019).*

[9] *Radford, Alec et al. “Language Models are Unsupervised Multitask Learners.” (2019).*

[10] D. V. Bagul and S. Barve, "A novel content-based recommendation approach based on LDA topic modeling for literature recommendation," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 954-961, doi: 10.1109/ICICT50816.2021.9358561.