AN INTELLIGENT TRAVEL RECOMMENDER SYSTEM BY MINING BEHAVIORAL ATTRIBUTES FROM ONLINE TRAVELOGUES IN MALAYALAM – A LOW RESOURCED LANGUAGE

Muneer V.K[1]*, Mohamed Basheer K.P[2]

Article History: **Received: 28.03.2023** **Revised: 20.04.2023** **Accepted: 06.05.2023**

Abstract

Language technology involves various language processing tools and techniques which significantly contribute to Natural Language Processing (NLP). Among NLP, natural language text and speech processing are two emerging segments that require huge attention from research. Regional language processing with the advent of Artificial Intelligence brings umpteen opportunities, especially in the Indian context as many languages were spoken in different parts of the Country. A Recommender Model in the Malayalam language in Travel and tourism domain using unsupervised machine learning techniques is the intention behind this paper. Malayalam is a low-resource and highly inflected language that possesses a greater chance for ambiguity. Data sharing online platforms and social media are used as data collection sources, where the availability is still limited and challenging, which may cause scarcity of data. The works propose various methodologies to generate a custom-made scraping model from the social media written in the Malayalam Language and its preprocessing. A deep-level Travelogue Tagger has been specially constructed as part of the experiment. This paper proposes a recommender model based on traveler reviews using Collaborative filtering and Cosine similarity methods. The experiment succeeded with high precision.

Keywords: Natural Language Processing, Recommender System, Unsupervised Learning, Part of Travelogue Tagger

[1*,2] Research Department of Computer Science, Sullamussalam Science College Areekode, University of Calicut, Kerala, India

***Correspondence Author:** Muneer V.K

*Assistant Professor, Department of Computer Science, Sullamussalam Science College Areekode, University of Calicut, India, vkmuneer@gmail.com

DOI: 10.48047/ecb/2023.12.si5a.0349

1. Introduction

For the past few years, Technology has been achieving growth at an exponential rate. People use it as an inexpensive mode of communication. Everybody connects with multiple applications through their mobile phones and handheld devices. Posting images, videos, review comments, polls, and other activities is not a big deal nowadays. That is the reason why Terabytes of data have been generated every second. According to a poll, social networks are used by roughly 73% of internet adults in the United States, and there are more than 2 billion accounts that produce a vast amount of online media content [1].

Information retrieved from social media is considered incredibly important for organizations, business firms, political leaders, advertising companies, and policymakers. This data is professionally processed to prepare future strategies, recognize business trends, and propose recommendations to customers. From the trajectory of applications in social networking sites, Facebook is considered bigger enough to be the third-largest country, next to China and India [2].

Gathering meaningful information from these giant data repositories is harder and more challenging in research and academia. The availability of exact data and its proper processing define the precision of the output. To automate data collecting, there are primarily three approaches. The first method uses web services, where the native site offers tools and technologies for legal data access. The second method uses a bespoke application to access API, and the third method uses coding to scrape [3] social media for raw content. Twitter provides two packages Tweepy and Twitter for retrieving information using python and R languages respectively [4].

Web scraping is the process of gathering data from the internet and saving it to any backend storage medium for later retrieval and analysis. Scraping can be carried out either manually, automatically, by software, or by a web crawler. The two steps of web scraping are obtaining the necessary resources from a chosen website using HTTP queries and extracting enough information from the site to be further parsed, reformatted, and organized into a structured style. With the use of tools and libraries like request, puppeteer, and selenium, the resources can be extracted in the first stage as HTML, XML, pictures, audio, or JSON forms. The packages like beautiful soup shall be used for parsing row data.

Facebook's API [5], which only allows for the retrieval of a limited selection of predetermined data in a predetermined way, making it difficult to gather all of the information from a user's postings, including their unique preference set. To overcome this, several free automatic web scraping solutions, such as Octoparse, Dexi.io, Outwit Hub, Scrapinghub, and Parsehub [6], are available to researchers. These tools can be used to collect real-time tweets and posts filtered by keywords, location, and language using packages like Rvest, R Selenium, rtweet, stream R, and R facebook. Additionally, user reactions like comments, likes, and shares from public Facebook pages [7] can also be collected. However, it is crucial to obtain users' consent before monitoring their private communications, including emails, posts on social media, or private conversations [8].

2. Literature Review

A study conducted by Md. Abu Kauser et al, about web crawlers. Web crawlers are automated computer programs that search and download internet data by moving across web pages [9]. The native APIs from facebook.com and Instagram help to retrieve data to the registered user through an App ID. App Secret [10] is an encryption mechanism to ensure the secrecy of data transfer. Anitha Ananthan discusses the methodologies to retrieve and analyse quality research papers published in Scopus, Springer, SCI and Wo S journals to generate a recommender system for the researchers [11]. Various techniques commonly adopted for information mining from social networking sites like FB and Twitter are discussed by Said. A, et al [12].

In paper [13] Ilham Safeek conducted a study to generate a suitable career path for customers from social media information like reviews, online activities, and reactions they share on online platforms. Utilizing data from social media, particularly tweets from Twitter, could potentially increase the accuracy of suggestions, according to research by Sunita Tiwari and colleagues [14]. Stefan and his colleagues concluded that there are four separate processes in the social media analytics process: data discovery, retrieving, planning, and investigation. [15].

The South Dravidian language Malayalam is an agglutinative language with a sophisticated inflectional morphology. The complexity of text processing in the Malayalam language is discussed by Rizwana in the paper [16]. Ajees A, et. al discussed a NER (Named Entity Recognition) model to low resourced language Malayalam using Neural Networks [17]. Hovy et al., classified the concept of text summarization as extractive and abstractive model [18]. Remiyya and her team conducted a study on retrieving relevant data from text written in the Malayalam language on social networking sites to extract different entities with the help of the Structured Skip-Gram Model [19]. A discussion was done by Pandian and the team on NLU (Natural Language Understanding) to understand input data as sentences using text or speech [20].

In a distinct document-level encoder built on BERT, Yang Liu et al. were able to articulate the document's general meaning, determine the meanings of its phrases, and achieve good results in both extractive and abstractive summarization [21] [22]. A graph-based approach to summarize the exact meaning of a passage was the study of Kanitha et al., in paper [23]. A semantic graph-based model and statistical sentence scoring algorithm proposed by Rajina and Sumam for text summarization [24].

3. Dataset

This study focuses on using data from a Facebook group called 'Sanchari', which is the biggest online travel group in the Malayalam language, with a member count of approximately 7.3 lakh as of 22-02-2022. The group provides a platform for individuals to share their travel experiences and reviews in the Malayalam language, covering different locations globally. With over 50000 detailed reviews and hundreds of comments per post, we considered 11500 travel posts from different users, each containing an average of 35 sentences. These posts provide personal preferences, public check-ins, likes, comments, and educational details, among others.

4. Methodology

This research paper discusses the limitations of using Facebook's API to fetch data from groups and pages. While the API provides a faster and more efficient method to collect data, it has limited availability and is insufficient for carrying out research using ML and NLP techniques for customized recommender models. As a result, a custom tool has been developed using JSON, Node JS, and other scripting tools to scrape essential details from Facebook. This tool allows for a more comprehensive collection of data and provides the necessary information for the development of a customized recommender model. The functionality of algorithms is based on three successive processes. Targeting the posts returns the posts, time of publishing, reactions, and comments received, shares count, and the author's profile_url for each post. For the time being, we didn't concentrate on responding to the post's comments. With the aid of this profile_url, the second stage concentrates more on the user's check-in information and public personal information like education, place, marital status, and family details. The final stage involves flattening the JSON format, which contains all these raw data, into an Excel sheet.

Facebook also offers group administrators an analytical tool called Insight, which will reveal thorough information about the group as an interactive dashboard. By observing statistical analysis, it is possible to assess user participation and find other indicators of group activities.

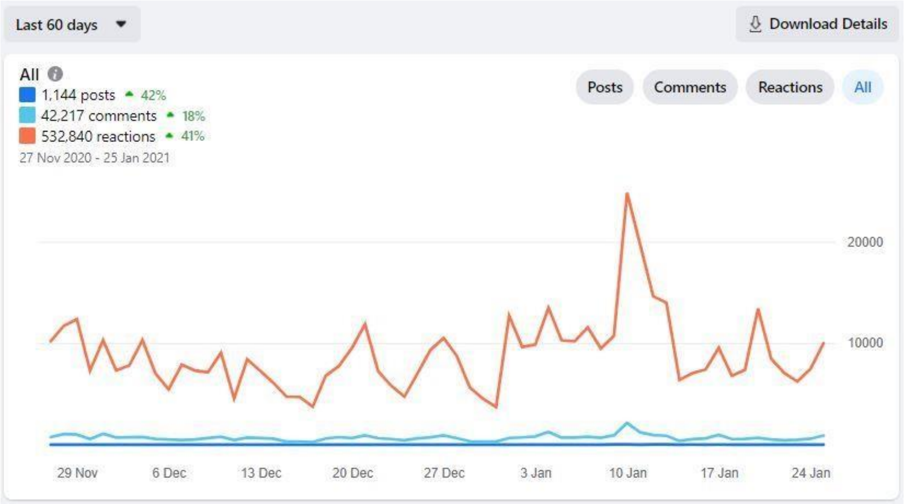


Fig. 1 Facebook group administrator insight

The steps followed in scraping FB posts from target Groups.

1. Sign into Facebook group/page/wall with username and password
2. Identifying and locating the exact DOM element from 'raw' data
3. Fetching user's posts, messages, posted time, and reactions.
4. Fetching individual favorites, histories, and location check-in details.
5. Creating and transforming JSON files to a spreadsheet.

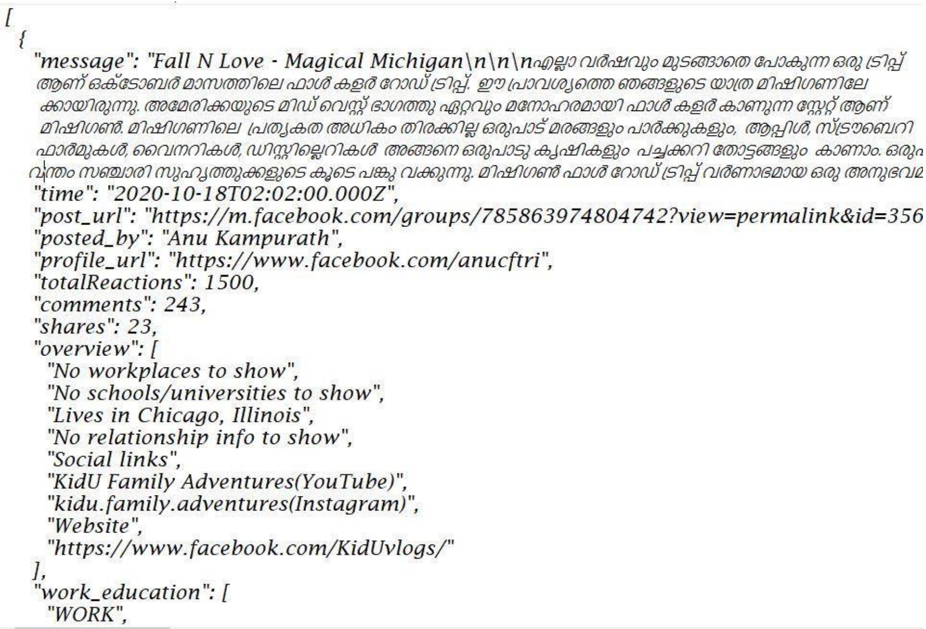
5. System Pipeline

The process of scraping is divided into two segments. In the first phase, it fetches all available details from the given Facebook group. The extracted data in the form HTML tags and JSON scripts are then converted into Excel spreadsheets or CSV files using three submodules. In the second phase, Natural Language Processing is used to preprocess Malayalam text from the Facebook group. The algorithm successfully extracted in the first phase and was able to preprocess it using NLP techniques to develop a recommender system model.

5.1. *Data collection from Facebook group*

Phase 1: Basic post scraping. An array is used here to append information from the travel posts and associated details. We can thoroughly scrape each post by iterating through this array. Native websites like Facebook and Instagram offer the natural Graph API, which may retrieve information much more quickly than specialized web scraping. But it can provide a pre-defined set of information only. That will not suit the requirement for this study. However, the system could get 2500 post details in an hour when using cutting-edge tools and optimized code. In total, 12500 posts with details were scraped in 5 hours while the code was running. To bypass rate restrictions and address memory leaks, the algorithm introduces random delays between each fetch, allowing for the retrieval of more posts. Since precise CSS selectors are not available on the page, the algorithm uses parent-child relations in the DOM. The data is then filtered and saved in JSON format for further processing in the next stage.

Phase 2: User's data fetching. The random time delays between each page visit in this step cause it to take longer than step one. Each post author visits 8 pages, with a 2 to 8-second pause between each page view. We determined that 2 to 8 seconds are the ideal range for these adjustable time intervals. Author information for a single post: 16 seconds minimum, or 2 times 8 seconds. Maximum duration: 64 seconds (8 * 8). On average, the details of one post's author are collected in one minute, even with delays during scrolling for scraping check-ins. 1500 or more posts were handled in a single day. By adjusting



```
[
  {
   "message": "Fall N Love - Magical Michigan\n\n\nഎല്ലാ വർഷവും മുടങ്ങാതെ പോകുന്ന ഒരു ട്രിപ്പ്
   ആണ് ഒക്ടോബർ മാസത്തിലെ ഫാൾ കളർ റോഡ് ട്രിപ്പ്. ഈ പ്രാവശ്യത്തെ ഞങ്ങളുടെ യാത്ര മിഷിഗണിലേ
   ക്കായിരുന്നു. അമേരിക്കയുടെ മിഡ് വെസ്റ്റ് ഭാഗത്തു ഏറ്റവും മനോഹരമായി ഫാൾ കളർ കാണുന്ന സ്റ്റേറ്റ് ആണ്
   മിഷിഗൺ. മിഷിഗണിലെ പ്രത്യകത അധികം തിരക്കില്ല ഒരുപാട് മരങ്ങളും പാർക്കുകളും, ആപ്പിൾ, സ്ട്രോബെറി
   ഫാർമുകൾ, വൈനറികൾ, ഡിസ്റ്റില്ലെറികൾ അങ്ങനെ ഒരുപാടു കൃഷികളും പച്ചക്കറി തോട്ടങ്ങളും കാണാം. ഒരു
  വ്ന്തം സഞ്ചാരി സുഹൃത്തുക്കളുടെ കൂടെ പങ്കു വക്കുന്നു. മിഷിഗൺ ഫാൾ റോഡ് ട്രിപ്പ് വർണാഭമായ ഒരു അനുഭവമ
   "time": "2020-10-18T02:02:00.000Z",
   "post_url": "https://m.facebook.com/groups/785863974804742?view=permalink&id=356
   "posted_by": "Anu Kampurath",
   "profile_url": "https://www.facebook.com/anucftri",
   "totalReactions": 1500,
   "comments": 243,
   "shares": 23,
   "overview": [
     "No workplaces to show",
     "No schools/universities to show",
     "Lives in Chicago, Illinois",
     "No relationship info to show",
     "Social links",
     "KidU Family Adventures(YouTube)",
     "kidu.family.adventures(Instagram)",
     "Website",
     "https://www.facebook.com/KidUvlogs/"
   ],
   "work_education": [
     "WORK",
```

Fig. 2. Post details fetched in step 1.

time delays and providing fast internet, these findings can be enhanced.

Parameters	Remarks	Cycle
Posts	Complete travel review	Cycle I
Time	Time of posting	Cycle I
post_url	Link of the post	Cycle I
profile_url	The distinctive ID of the user	Cycle I
total_reactions	Likes, comments, shares	Cycle I
Comments	User comments	Cycle I
shares	Shares count	Cycle I
about_work_ and_education	Job and education	Cycle II
about_family_ and_relation	Marital status	Cycle II
Check-ins	traveled places and check-ins	Cycle III

Table 1. Data scraping details in iteration wise.

5.2. *Language Processing over Malayalam travelogue*

The travelogues and messages were obtained from a Facebook group that has in-depth descriptions of various travel experiences in Malayalam. The paragraphs are to be subjected to perform sentence tokenization and these sentences to work tokenization. General preprocessing techniques such as removal of stop words, punctuations, symbols, images, numbers, and codemixed tokens. Another important task here to perform is to identify the root word of each token. Root_pack is a python package developed by ICFOSS for this purpose [25].

Implementation of the root_pack package is as below, import root_pack
root_pack.root("സഹോദരനോടൊപ്പമായിരുന്നു")
The output from this package is the root word "സഹോദരൻ"

5.3. *Look-up Dictionary Formation*

The lengthy travelogue undergoes basic data cleaning and text processing tasks. Sentence tokenization and word tokenization then lead to root word extraction. The next phase is to stop word removal and punctuation removal. Malicious and irrelevant tokens are omitted in this step. The next major task is annotating these refined tokens with corresponding tags.

The travelogue must be sorted with a predefined set of Travel tags. The most essential details must be identified from each travelogue are Travel Type, Mode of Transportation, Location details, Climate of Travel, and Geographical features of the destination. For that, a look-up dictionary has been created with the above tags. Each list contains all possible fields in that genre. The sample structure of the look-up dictionary is given below in Table 2.

Locations:		മുക്കം, മണാലി, ദൽഹി, കശ്മീർ, മലന്പുഴ, etc
Travel Types:	[നടത്തം]	നടത്തം, ലിഫ്റ്റ്
	[റോഡ്]	കാർ, ബസ്, ഡ്രൈവ്, റോഡ്.
Travel Modes:	[സോളോ]	സോളോ, തനിച്ച്
	[കൂട്ടുകാർ]	കൂട്ടുകാർ, കുടുംബം, ചങ്ക്, ബ്രോ
Climates:	[തണുപ്പ്]	തണുപ്പ്, മഴ, ശൈത്യം
	[ചൂട്]	ചൂട്, സമ്മർ, വെയിൽ, വേനൽ

Table 2. Tabular format of look-up dictionary.

Once the travelogue has been processed with the above steps, each relevant token in Malayalam travelogue has been annotated with a corresponding tag and embedded in numerical values. Thus, each travelogue can be defined into four or five discrete features which would be sufficient to describe that post.

5.4. *Malayalam Travelogue Tagger*

To fetch and annotate all tokens such as Destination, Mode of journey, Type of journey, and time of travel in the travelogue, the existing POS Tagger for Malayalam has been updated by appending additional tags for dealing Travelogues. The Part of Travelogue (POT) Tagger [34] follows the BIS of the POS Tagger. The structure of POT Tagger has been given below:

The additional tags added to POT Tagger are,
Place/Location (L): ഡൽഹി, ലണ്ടൻ, മണാലി / Delhi, London, Manali
Type of Travel (TT): കാർ, ട്രെയിൻ, റോഡ്, ബൈക്ക് / Car, Train, Road, Bike
Mode of Travel (TM): ഒറ്റക്ക്, സഹപ്രവർത്തകർ, കുടുംബം, കൂട്ടുകാർ / Solo, Colleagues, Family, Friends
Climate of destination (LC): തണുപ്പ്, മഞ്ഞ്, സമ്മർ / Hot, Summer, winter, snow
Type of destination (LT): പ്രകൃതി, തീർത്ഥാടനം, ചരിത്രം / Nature, Pilgrimage, High range

The processing of tagging and other language processing is done with the help of Natural Language Toolkit (NLTK) packages and custom paid packages and Corpus files in the python programming language. Sample POT Tagger is given in Table 3.

Sl. No	Category		Label	Annotation Convention	Example
	Top Level	Sub Level			
I	Location	1	L	L_N	ഡൽഹി, മുന്നാർ
II	Type of trip	1	TT	TT_T	കാർ, ബോട്ട്, വിമാനം
III	Mode of trip	1	TM	TM_N	ചങ്ങാതി, ഒറ്റക്ക്, കുടുംബം
IV	Destination Climate	1	LC	LC_T	മഴ, ചൂട്
V	destination Type	1	LT	LT_N	സാഹസികം, ചരിത്രം

Table 3. Customized tags and annotations are used in POT Tagger.

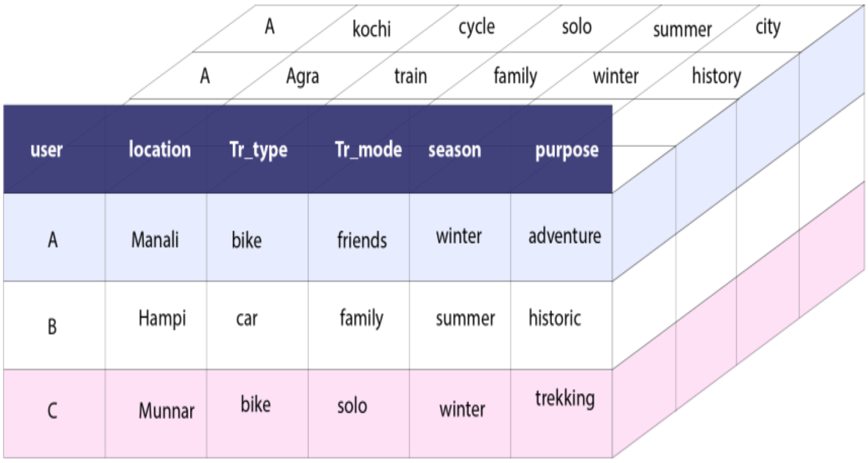
5.5. *Preparation of Travel DNA*

Travel DNA gives the choices and preferences of each traveler. Each travelogue is now able to describe its crux in the 5 most essential features. A traveler may visit different locations in different modes and different travel types. The taste of a particular traveler can be figured out from observing these records and can create a pattern for his travel. Such travel patterns of every traveler can form Travel DNA and be further classified into different clusters, where each cluster contains like-minded travelers as shown in Table 4.

Traveler	Location	Tr__Type	Tr_Mode	Climate	Location Type
Babu	Delhi 1	Bus 1	Family 3	Rain 1	Historic 1
	Mumbai 3	Car 2	Friend 3	Winter 1	Pilgrimage 3
	Goa 4	Flight 2	Friend 2	Winter 1	Adventure 5
James	ആഗ്ര 5	ട്രെയിൻ 4	കുടുംബം 3	തണുപ്പ് 1	ചരിത്രം 2
	ഗോവ 2	കാർ 2	ഒറ്റക്ക് 1	മഞ്ഞ്	വിനോദം 8
	കൊച്ചി 9	സൈക്കിൾ 6	സുഹൃത്ത് 3	തണുപ്പ് 1	തീർത്ഥാടനം 3
	ഗോവ 25	വിമാനം 5	സുഹൃത്ത് 3	സമ്മർ 2	വിനോദം 8

Table 4. Sample the Travel DNA of users with 5 distinct keywords from each travelogue.

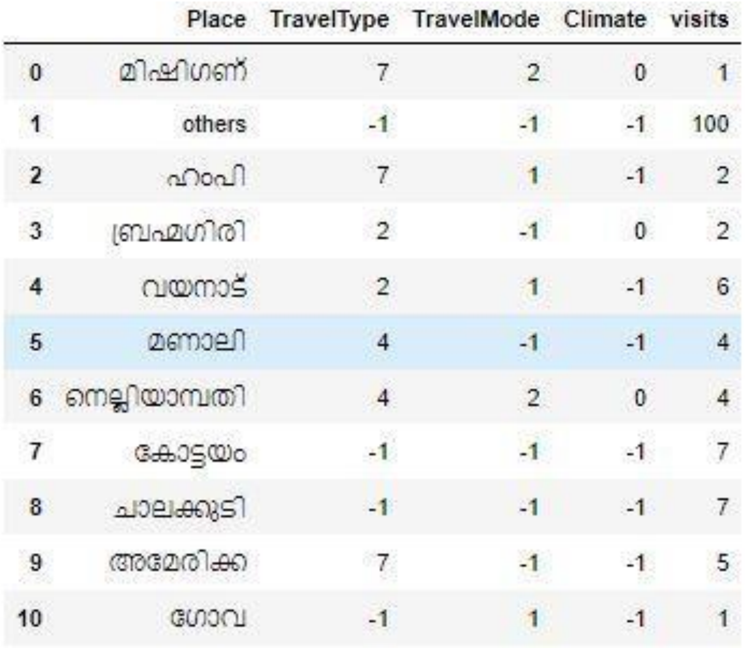
Travel DNA is good enough to provide comprehensive information about the travel pattern of each traveler. The unique feature set for each traveler is created by observing the most frequent mode of travel or destinations. The DNA of A, B and C are symbolically given as a 3D representation in Figure 4. Travel A has traveled multiple locations with multiple combinations of features. Likewise, DNA of each traveler constructed for further processing.



user	location	Tr_type	Tr_mode	season	purpose
A	Manali	bike	friends	winter	adventure
B	Hampi	car	family	summer	historic
C	Munnar	bike	solo	winter	trekking

Fig. 4. A Sample representation of Travel DNA of 3 users.

A python package count_vectorizer() is used for this purpose. As the Travel DNA is created, the next process is to create Location DNA. This is also an informative table that contains location details, such as the mode (TM) in which, with whom most travelers preferred to reach here, the type (TT) in which type of vehicle most travelers opted and the location climate (LC) in which season people select this destination. A sample structure of Location DNA is given in Figure 5.



	Place	TravelType	TravelMode	Climate	visits
0	മിഷിഗൺ	7	2	0	1
1	others	-1	-1	-1	100
2	ഹംപി	7	1	-1	2
3	ബ്രഹ്മഗിരി	2	-1	0	2
4	വയനാട്	2	1	-1	6
5	മണാലി	4	-1	-1	4
6	നെല്ലിയാമ്പതി	4	2	0	4
7	കോട്ടയം	-1	-1	-1	7
8	ചാലക്കുടി	-1	-1	-1	7
9	അമേരിക്ക	7	-1	-1	5
10	ഗോവ	-1	1	-1	1

Fig. 5. Summarized list of Location DNA

6. Experimental Results

After constructing these two structures, the cosine similarity method is used to identify the list of suggested places for each user. Cosine similarity is a widely used metric in natural language processing (NLP) to quantify the similarity between two vectors. In the context of this study, the cosine similarity is calculated by measuring the angle between two vectors that represent the locations of different travel destinations in a multi-dimensional space as given in Figure 6. As the mathematical calculation given in Figure 7, the similarity between two travel destinations enables the algorithm to prepare a list of recommended destinations for new users. The RS is designed to generate two level of predictions. In primary list of recommendation displays the places user never visited and is likely to visit next. The secondary list of recommendations contains a set of places where like-minded users visited and hence the current user may prefer to visit.

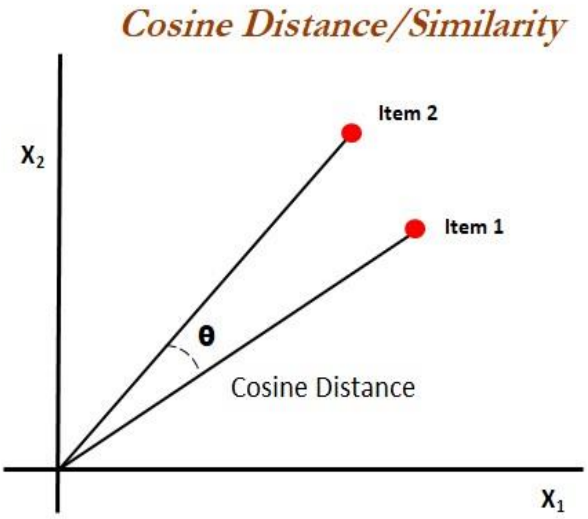


Fig. 6. Representation of cosine distance between two vectors.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2} \sqrt{\sum\limits_{i=1}^{n} B_i^2}},$$

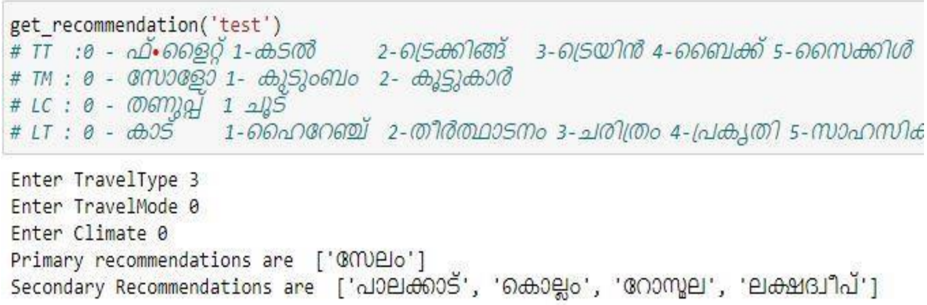
Fig. 7. Mathematical equation of Cosine similarity

In a manual evaluation, from the given user input, the model could generate both primary and secondary lists of destinations. Out of 5 recommendations, 4 of them are correct in the secondary and 3 of them are correct in the primary list, as an average which means 80% accurate in secondary and 60% in primary recommendations.

User	TT	TM	LC	Recommendations	
				Primary	Secondary
Bibin joseph	0 റൈഡ്	0 സഹോദരൻ	0 തണുപ്പ്	റാണിപുരം, ഓമശ്ശേരി	മൂന്നാർ, വയനാട്
	3 ട്രെയിൻ	1 കുടുംബം	0 തണുപ്പ്	പൊള്ളാച്ചി, ലഡാക്ക്	കോട്ടയം, പഞ്ചാബ്
	0 റൈഡ്	1 കുടുംബം	1 മഞ്ഞ്	വയനാട്	കാഞ്ഞങ്ങാട്, ഇടുക്കി
Test	2 ട്രെക്കിംഗ്	1 കുടുംബം	0 തണുപ്പ്	രാജസ്ഥാൻ	നിലന്പൂർ, മേഘാലയ
	3 ട്രെയിൻ	0 സഹോദരൻ	0 തണുപ്പ്	സേലം	കൊല്ലം, ലക്ഷദ്വീപ്
	0 റൈഡ്	0 സഹോദരൻ	1 മഞ്ഞ്	ഗോവ	ഗവി, തെന്മല

Table 5. Different combinations of features for users and their predictions

From Table 5, the different recommendations in both the primary list and the second list have been predicted for various users. Depending upon the change in the combination of features recommendation list has been updated accordingly. A sample output list is given in Fig. 8.



```
get_recommendation('test')
# TT  :0 - ഫ്ളൈറ്റ് 1-കടൽ      2-ട്രെക്കിങ്ങ്   3-ട്രെയിൻ 4-ബൈക്ക് 5-സൈക്കിൾ
# TM : 0 - സോളോ 1- കുടുംബം  2- കൂട്ടുകാർ
# LC : 0 - തണുപ്പ്  1 ചൂട്
# LT : 0 - കാട്     1-ഹൈറേഞ്ച്  2-തീർത്ഥാടനം 3-ചരിത്രം 4-പ്രകൃതി 5-സാഹസിക

Enter TravelType 3
Enter TravelMode 0
Enter Climate 0
Primary recommendations are  ['സേലം']
Secondary Recommendations are  ['പാലക്കാട്', 'കൊല്ലം', 'റോസ്മല', 'ലക്ഷദ്വീപ്']
```

Fig. 8. Summarized A sample Recommendation for a Guest(test) user.

A model can be built by following this algorithm with the help of a large dataset. For this purpose, we trained the model with 11500 lengthy travelogues. With the help of Collaborative filtering and cosine similarity, the model can recommend the most suitable 5 locations for a guest customer or existing traveler.

7. Conclusion

This research project focuses on the development of a personalized tourist spot recommender model by developing a data scraping algorithm for travel-related posts from online repositories, especially from the largest Facebook Travel group in Malayalam. Social media platforms like Facebook offer a vast source of data, including travelogues, photos, check-ins, polls, opinions, activities, and updates, making it an ideal resource for data mining. Our algorithm aims to overcome the limitations of existing traditional data scraping methods. For this study, we gathered comprehensive information from 11,500 Facebook postings, 3781 public profiles of travelers, and details on 84463 location check-ins worldwide. Additionally, we developed a Part of Travelogue (POT) Tagger, a customized annotator for mapping travelrelated tokens, to prepare the Malayalam travel dataset. Travel DNA and Location DNA have been constructed with the help of this POT Tagger. The user's posts were further processed using a collaborative filtering and cosine similarity approach to provide them with a recommender model powered by machine learning. The model could suggest the most suitable 5 destinations as Primary recommendations for users considering the travel histories, choices, and preferences combinations. The performance of the algorithm can be increased by considering additional features like age group, and gender, optimizing POT Tagger, and enhancing tokens in the corpus. Unsupervised clustering algorithms like K-means clustering or hybrid agglomerative algorithm shall be considered for better clustering and prediction. The significance of algorithm is as it is the first Personalized Travel recommender model developed for the Malayalam language.

References

1. Amanda Lenhart, Kristen Purcell, Aaron Smith, Kathryn Zickuhr, (2012), Social Media & Mobile Internet Use Among Teens and Young Adults, 2010.
2. Dr. M. Saravana kumar, Dr. T. Sugantha Lakshmi, Social Media Marketing, Life Sci Journal, (2012);9(4): 4444-4451. (ISSN: 1097- 8135)
3. Bogdan Batrinca, Philip C. Treleaven, Social media analytics: a survey of techniques, tools and platforms, 2014, Springer, AI & Soc (2015) 30:89–116, DOI 10.1007/s00146-014-0549-4.
4. Kuldeep Singh, Harish Kumar Shakya, Bhaskar Biswas. Clustering of people in social network based on textual similarity Perspectives in Science (2016) 8. 570—5732016, http://dx.doi.org/10.1016/j.pisc.2016.06.023.
5. Jan-Willem van Dam, Michel van de Velden (2015), Online profiling and clustering of Facebook users, Decision Support Systems 70 (2015) 60–72, http://dx.doi.org/10.1016/j.dss.2014.12.001 0167-9236/© 2014.
6. Plamen Milev (2017), Conceptual Approach for Development of Web Scraping Application for Tracking Information, Economic Alternatives, 2017, Issue 3, pp. 475-485
7. Bernhard Rieder (2013), Studying Facebook via Data Extraction, WebSci '13, Proceedings of the 5th Annual ACM Web Science Conference, 2013, Pages 346–355. https://doi.org/10.1145/2464464.2464475.
8. Sean C. Rifea, et al,(2014), Participant recruitment and data collection through Facebook: the role of personality factors, International Journal of Social Research Methodology, 2014, http://dx.doi.org/10.1080/13645579.2014.957069.

9. Md. Abu Kausar (2013), V. S. Dhaka , Sanjeev Kumar Singh, Web Crawler: A Review, International Journal of Computer Applications, 2013, Volume 63– No.2, February 2013, DOI: 10.5120/10440-5125.
10. Lusiana Citra Dewi, Meiliana, Alvin Chandra, (2019) Social Media Web Scraping using Social Media Developers API and Regex, Procedia Computer Science 157 (2019) 444–449.
11. ANITHA ANANDHAN, et al, (2018) Social Media Recommender Systems: Review and Open Research Issues, 2018, 2169-3536 2018 IEEE. Translations, DOI 10.1109/ACCESS.2018.2810062.
12. Said A. Salloum, Mostafa Al-Emran, Azza Abdel Monem, Khaled Shaalan, (2017) A Survey of Text Mining in Social Media: Facebook and Twitter Perspectives, Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 1, 127-133 (2017)
13. Ilham Safeek, Muhammad Rifthy Kalideen, (2017) PREPROCESSING ON FACEBOOK DATA FOR SENTIMENT ANALYSIS, Proceedings of 7th International Symposium, SEUSL, 7th & 8th December 2017.
14. Sunita Tiwari, et. al, (2019) Implicit preference Discovery for biography Recommender system Using Twitter, International Conference on Computational Intelligence and Data Science, 2019, DOI: 10.1016/j.procs.2020.03.352.
15. Stefan Stieglitz et.al, [2018] Social media analytics – Challenges in topic discovery, data collection, and data preparation, International Journal of Information Management 39 (2018) 156–168, https://doi.org/10.1016/j.ijinfomgt.2017.12.002.
16. Thandil, R.K., Basheer, K.P.M. (2022). Speaker Independent Accent Based Speech Recognition for Malayalam Isolated Words: An LSTM-RNN Approach. In: Dev, A., Agrawal, S.S., Sharma, A. (eds) Artificial Intelligence and Speech Technology. AIST 2021. Communications in Computer and Information Science, vol 1546. Springer, Cham. https://doi.org/10.1007/978-3-030-95711-7_2
17. Ajees A Pa, , Sumam Mary Idicula, (2018) A Named Entity Recognition System for Malayalam using Neural Networks, 2018, Procedia Computer Science 143 (2018) 962–969.
18. Eduard Hovy, Chin-Yew Lin, Automated Text Summarization in SUMMARIST. In Advances in Automatic Text Summarization, I. Mani and M. Maybury (editors), 1999.
19. Remmiya Devi G, et.al, (2022) Entity Extraction for Malayalam Social Media Text using Structured Skip-gram based Embedding Features from Unlabeled Data, 2016, doi: 10.1016/j.procs.2016.07.276, Procedia Computer Science 93 (2016) 547 – 553.
20. Pandian, S.. (2019). Natural Language Understanding of Malayalam Language. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING. 7. 133-138. 10.26438/ijcse/v7si8.133138
21. Yang Liu and Mirella Lapata, (2019) Text Summarization with Pretrained Encoders, arXiv:1908.08345v2 [cs.CL] 5 Sep 2019.
22. Derek Miller, Leveraging BERT for Extractive Text Summarization on Lectures, 2019, arXiv:1906.04165.
23. Kanitha D K, et al, Malayalam Text Summarization Using Graph Based Method, International Journal of Computer Science and Information Technologies, Vol. 9 (2) , 2018, 40-44, ISSN:0975-9646.
24. Rajina Kabeer and Sumam Mary Idicula, (2014), Text Summarization for Malayalam Documents – an Experience, 2014, International Conference on Data Science & Engineering (ICDSE), 978-1-4799-5461-2114/$31.00 @2014 IEEE. [25] “root-pack,” PyPI, Jun. 21, 2019. https://pypi.org/project/root-pack/