International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

Trip Planner: A Big Data Analytics Based Recommendation System for Tourism Planning

Kamal Kumar Ranga[1], Chander Kumar Nagpal[2], Vedpal[3]
[1]Department of Computer Engineering,
JC Bose University of Science and Technology,
Faridabad, Haryana, India
*kamal.ranga@gmail.com
[2]Department of Computer Engineering,
JC Bose University of Science and Technology,
Faridabad, Haryana, India
nagpalckumar@rediffmail.com
[3]Department of Computer Engineering,
JC Bose University of Science and Technology,
Faridabad, Haryana, India
ved_ymca@yahoo.co.in

Abstract- Foreign tourism has gained immense popularity in the recent past. To make a rational decision about the destination to be visited one has to go through variety of social media sources with very large number of reviews, which is a tedious task. Automated analysis of these reviews is quite complex as it involves non structured text data having slang terms also. Moreover, these reviews are pouring in continuously. To overcome this problem, this paper provides a Big Data analytics-based framework to make appropriate selection of the destination on the basis of automated analysis of social media contents based upon the adaptation and augmentation of various tools and technologies. The framework has been implemented using Apache Spark and Bidirectional Encoder Representation Transformers (BERT) deep learning models through which raw text review are analysed and a final score based on five metrics is obtained to recommend destination for visit.

Keywords-Big Data Analytics, Sentiment Analysis, Sentiment Score, International Tourism, Recommendation System, Deep Learning.

I. INTRODUCTION

The people around the world are fascinated to know about the cultures of different countries, lifestyles, food verities, monuments and historical places etc. creating the allure for global tourism [1]. Recent past decades have shown a major inclination for the foreign travel as there has been a significant increase in foreign travel frequency by Indians. This increase in number of travellers is from 4.42 million in 2000 to 26.9 million in 2019 though there has been a decrement to 7.29 million in 2020 due to pandemic [2]. This is due to increase in per capita income (from $443 per annum in 2000 to $2101 in 2020) there by significant rise in the size of middle and upper class of the country[3].Since this market has grown up like anything, many facilitators such as make_my_trip[4], TripAdvisor[5],ease_my_trip[6] and yatra [7] have come up to capture the business opportunities in this growing market.

Conventional mechanism to select a tourism destination involves reading and analysis of reviews available on various websites such as make_my_trip [4], TripAdvisor [5] etc. and the social media platforms like twitter [8], Facebook (meta)[9], Instagram [10] and YouTube[11] etc. Lots of travellers' bloggers post their videos having detailed information in terms of the best hotels, weather, famous sites to visit, transport-related reviews etc. for the information of prospective travellers. Most of the countries whose economy is dependent upon the tourism also provide the tourism related information on their respective websites. A person willing to travel abroad can refer to all these information for making the travel for his/her family convenient.

However, referring to all the above-mentioned information involves many problems: Posted text reviews are lengthy and their number is quite huge. Reading all these reviews and analysing them is quite a herculean task. Compared to the volume of the text, the information content is quite low. The information is spread over multiple websites and exploring it in totality is not feasible.

Thus, there is a need for an automated mechanism that can provide gist of these text reviews which are continuously growing and updating. If the analysis mechanism can also provide the information on the desired features (as per user requirement) then it will be an added advantage. The proposed work is an effort in this direction.

The work proposes a Big Data based framework that revolutionizes the process of selection of destination by

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

extracting and analysing the reviews posted in the form of natural language on different sources like TripAdvisor, YouTube and Twitter etc. The analysis of results is presented as a score for the place. The five common factors have been used to calculate the score of a place based on the reviews which are: availability of vegetarian food, facilities in hotel rooms, comfort, sightseeing & attractions and family friendliness. These factors are of most common concern for an individual person or family at the time of the selection of a destination to visit. The proposed framework is based on the data extracted from Top 10 destinations [12] where the Indian tourist have visited most in recent years. These places are Bali, Bangkok, Dubai, London, Mauritius, New York, Paris, Singapore, Sri Lanka, Switzerland.

In the proposed framework, text data is retrieved from the trusted sources and filtered using Selenium [13][14] and BeautifulSoup [15][16]. Filtering is performed to obtain reviews of the restaurants containing veg food, hotels which are suitable for family, places having common facilities, attractions and sightseeing at a place and comfort at the place. The data so obtained is quite voluminous, has variety/heterogeneity, high incoming velocity etc. making it a big data centric problem. The Apache Spark [17] (Big Data framework) is used to handle the huge amount of input data. The data obtained from Apache spark is fed to Natural Language Toolkit (NLTK)[18] for the purpose of data cleansing and pre-processing .The pre-processed data is provided to pre-trained BERT[19] for the purpose of sentiment analysis. The proposed framework converts the results into an elegant user-friendly visual representation. The user can see a visually attractive globe having all the information related to the destination with scores of different factors in dynamic format. The proposed framework helps the user not only in searching destination(s) according to his/her preferences but also as per the appropriateness for his/her family.

II. ORGANIZATION OF THE PAPER

The paper has been organized as follows: Section 2 takes up the literature survey in which the currently available work in the related domain has been discussed. Section 3 provides the guidelines and scope for the framework of proposed recommendation system in the light of literature review. Section 4 contains the basic details of the proposed system using Big Data analytics as the base. Section 5 discusses the phase-wise implementation details. Section 6 provides the experimental set up details and design of metric. Section 7 presents the result obtained and analysis thereof. The results have been presented on the world map for clear and better understanding. The paper ends with conclusion and future scope of the proposed work.

III. LITERATURE REVIEW

Before proceeding for the proposed work, a comprehensive study was made to understand and assess the available work in the domain of tourism recommendation system based upon user reviews on various social media platforms. An overview of the available work is as follows:

Ana Reyes-Menendez et al. [20] have emphasized the emerging importance of electronic word of mouth (e-WOM) in customer decision-making, specifically in the tourism sector. The Elaboration Likelihood Model (ELM) has been used for analysis with five main factors namely: Volume of e-WOM, Source, Rating, Customer Participation and Perceived e-WOM Credibility. They claimed that the strategy is helpful for hotel and tourism platforms managers in devising strategies to improve their online reputation.

S. M. Al-Ghuribi et al. [21] applied aspect-based sentiment analysis (ABSA) on small-sized labelled datasets. They are of the view that the real datasets such as TripAdvisor contain gigantic reviews which require huge computing resources and provide spurious results due to ambiguity of natural language.

K. S. Ntalianis et al. [22] presented a geographical locations rating scheme based on the crowd-sensing and crowd-sourcing achieved via smartphone applications and websites. Evaluation is performed by using the six feelings and five strength levels. The experimented Results of real-world data are compared with Google Maps and TripAdvisor rating shows the efficacy of the proposal.

Benlahbib et al. [23] proposed a new approach that combines different attributes of review like helpfulness, time, rating and sentiment to give a single value reputation index. The approach works in in three phases. In first phase BERT model is used to fine-tune the sentiment orientation. In second phase, a numerical score to each review is assigned and aggregated. In last phase obtained results are visualized as reputation value and opinion categories.

S. M. Al-Ghuribiet al. [24] have proposed a multi-criteria recommender system (MCRSs) to improve accuracy of Recommender System(RS) performance by incorporating review elements relevant to entire system or the specific feature.

I. Topal et al. [25] have applied artificial intelligence to analyse the data from TripAdvisor on the basis of travel history as well as travel preferences. As a result of their work, travel potential for Turkey was identified and 2018 was declared as “Turkey Tourism Year” in China.

K. A. Fararni et al. [26] claims that tourist data has burgeoned multi-fold with the development of Online Travel Agencies (OTA). They proposed a conceptual framework for tourism recommender system based on a hybrid approach using Big Data technologies, Artificial Intelligence and Operational

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

Research to promote tourism in Morocco, specifically in the Daraa-Tafilalet region.

G. Adomavicius et al. [27] have provided an overview and description of current generation recommender systems, their classification, limitations and also discusses possible extensions that can improve their capabilities. These extensions include improving user understanding, incorporation of contextual information and multi-criteria ratings.

G. Sun et al. [28] have presented a visualization technique called route-zooming that can embed spatio-temporal information into a map seamlessly for occlusion-free visualization of both spatial and temporal data.

J. He et. al. [29] has proposed multimodal sentiment analysis that focuses on language, acoustic and visual information with BERT as base technology. An internal updating mechanism has also been proposed to avoid the over-fitting of the model in the training process.

H. C. M. Senefonte et al. [30] proposed PredicTour an approach that processes user check-ins on location-based social networks (LBSNs) to predict movement patterns with or without previous visiting history.

P. Nitu et al. [31] proposed a social media analytics-based travel recommendation system that recommends place of interest based on user's need and preferences. For this twitter data of users, their friends and followers is analysed to obtain insights of recent travel interest.

In addition to the above-mentioned papers, other papers [32-38] have also been studied during the literature survey. After going through the literature following observations were made.

The social media platforms can provide the requisite data in the form of user reviews for analysing tourists' preferences. Moreover, this data is quite recent and more relevant compared to the other available sources such as news articles, magazines etc.

- The data available on the social media platforms is mostly in form of text review which can be analysed for the purpose of sentiment analysis by using the tools like BERT after proper training.
- The data available in the form of videos and images requires huge storage and other resources and is far more complex to process.
- The data available on the social media platforms is quite huge and continuously updating, therefore the tool applicable to Big Data analytics can be quite useful.
- The use of the maps to show the location along with its tourists' preference indicators can be quite meaningful, informative and eye catching.
- The available work in the domain doesn't takes up the problem in the holistic manner and tries to solve the various aspects in individual / isolated manner.

The above-mentioned observations helped us in deciding the scope and guidelines of the proposed work which follow in the next section.

IV. Scope and Guidelines of Proposed Work

The proposed work in intended to take up the following guidelines to meet its scope:

- There should be a single unified automated mechanism for extraction and modelling of incoming data so as to speed up the process.
- Normally, the extracted data is collected into huge number of text files. Handling and processing such large number of makes the system complex that makes it beyond the capability of normal computer system. Thus a specialized platform is required to come out of this problem.
- The proposed system should be totally automated and integrated to solve the problem in holistic manner.
- Instead of qualitative result, the system should provide quantitative results with the help of suitably designed metric(s).

The subsequent section provides the details of proposed recommendation system framework.

V. Proposed Recommendation System Framework

As already described in the introduction section, proposed Recommendation System framework has been designed for tourism planning using Big Data tools. The designed framework involves six phases as shown in Fig.1.First phase involves the formal definition of the problem and statement of objectives for the work to be carried out. This includes the identification of goals to be achieved and the description of its scope and limitations. The required results and their format are normally finalized, to a major extent, at the time of the statement of objectives yet they may evolve as the project processes. It helps in making a decision about the actual processing scenario. The relevant and credible data sources are identified in second phase. The identified source must comply with the basic requirements of accuracy, completeness, reliability, relevance and timeliness etc. Third phase comprises the collection of the relevant data from the identified sources. The collected data may be stored in a temporary storage or permanent storage as per the requirement of the problem solution.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

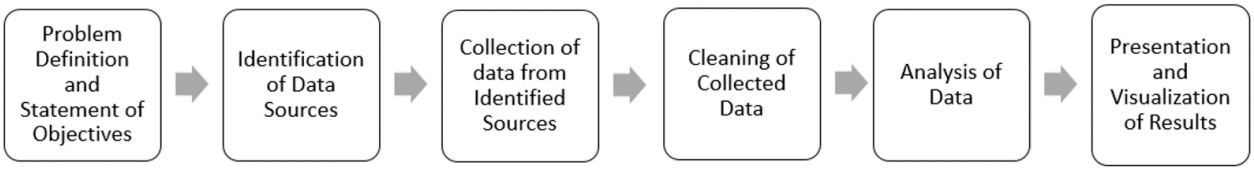

Fig.1 Framework of the Proposed Big Data Analytics based Recommendation System.

In fourth phase, cleansing of the data is performed as per the requirement of the problem. Data cleansing makes sure that the data reflects reality by removing duplicates, filling of missing values and the elimination of typographical errors etc. Fifth phase transforms data into requisite formats for the purpose of analytics. Properly formatted, cleansed and validated data improves data quality and protects applications from potential landmines such as null values, unexpected duplicates, incorrect indexing etc.[39]. The meaningful results in required format are obtained by actually processing the data. Sixth and final phase presents the results keeping in view the user requirement, aesthetics and effectiveness of presentation.

VI. PHASE WISE IMPLEMENTATION DETAILS

After describing the overview of the recommendation system, let take up phase wise implementation details as used in our proposed work.

A. Phase 1: Problem Definition and Statements of Objectives

- *Problem Definition*

To identify the suitable tourist destination(s) for the Indian families, keeping in view the availability of vegetarian food, family friendliness, comfort, sightseeing and other general facilities based on the social media analytics.

- *Statement of Objectives*

a. To identify relevant and reliable social media sources for data collection.
b. To identify the various tools for efficient data extraction.
c. To cleanse and reformat the data to make it suitable for analysis.
d. To devise a suitable mechanism for analysing textual data for obtaining positive and negative sentiment scores.
e. To develop a suitable index/ metric to represent the consolidated relative intensity of positive and negative sentiments.
f. To create a suitable format for representing the obtained results, keeping in view the ease of understandability, aesthetics and meaningfulness.

B. Phase 2: Identification of Data Sources

As per the stated objectives, various data sources were explored to capture the relevant data. List of explored data sources is shown in Table 1.

Table 1: List of Explored Data Sources

Available Data Source	**Type of Data**
Twitter	Text Tweets, Images, Videos
YouTube	Text Comments, Videos
TripAdvisor	Text Comments, Images, Videos
MakeMyTrip	Text Comments, Images, Videos
Booking.com	Text Comments, Images, Videos
Trivago	Text Comments, Images, Videos
Facebook	Text Comments, Images, Videos

The data available in explored sources is in the form of text, images and videos. The automated analysis for images and video data is quite cumbersome task and leads to inefficient analytical results. Moreover, time and storage space requirements are also very high as compared to text data. Keeping this aspect in view, only textual data from the identified sources has been considered. Amongst the various data sources listed in Table 1, only three data sources, namely TripAdvisor.com, Twitter and YouTube, have been considered for collecting the relevant data. The reason for selecting these specific data sources is given below:

- *TripAdvisor*

TripAdvisor.com is one of the world's most trusted travel research platforms that allows its users to post reviews and opinions on its platform. It aggregates the user's feeling about the hotels, price, destinations, stay, restaurants and other such activities throughout the world through its flagship

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

TripAdvisor brand. Reasons for choosing TripAdvisor are as follows:

1. A wide user base across the world.
2. Only verified users have the right to post reviews.
3. Has a dedicated team to verify the authenticity of reviews posted by users.
4. No anonymous user can post a review or comment on the platform.

- *Twitter*

Twitter is a micro blogging site that lets users share short posts called tweets. Reasons for selecting Twitter are:

1. Tweeted text is small (upto 280 characters) having relevant text only.
2. Several hashtags are decided by the authors. These hashtags can be used to identify the domain related to the tweets making it convenient to identify and scrap relevant data.
3. Availability of built-in API for data scrapping.

- *YouTube*

YouTube is one of the world's most popular and heavily used video sharing sites, where users can watch or upload their videos. The users can like, comment or share them too. From YouTube only comments (Textual data) from relevant video are extracted. The reasons for choosing YouTube as data source are:

1. Most of the travel videos on YouTube are associated with a large number of textual comments expressing the opinions of different people.
2.The opinions under such videos are normally very relevant and can be taken into consideration.

C. Phase 3: Collection of data from Identified Sources

Data Scrapping: The process of extracting desired information from a website or any other internet source is termed Data Scraping or Web Scrapping. It is used to extract unstructured information from a internet data source, which can be stored in a semi-structured format. In this paper, a web crawler has been implemented using python3 to extract user's review from chosen data sources. Selenium has also been used as a web automation bot that extracts data. It is capable of sending standard python commands irrespective of variation in browsers. Selenium alone would not be enough for the task of scrapping; it requires support of an open-source python library named BeautifulSoup to pull out information from webpages. The BeautifulSoup extracts required information from sections of the webpage; remove the markups, links and titles and saves only text data into storage. The combination of selenium and BeautifulSoup provides an efficient way to extract the relevant reviews. The snapshot of extracted data stored in a CSV file is shown in Table 2.

Table 2. A snapshot of Extracted Raw Data.

title	text	date_of_stay	helpful_votes
Highlight of my trip!	You will simply go over to the top of the rice terrace and simply see the green terraces (or I would say just green grasses). No explanation of how rice / pady is prepared. No demo of any manufacturing activity. Infact the hot weather will further frustrate you	Date of experience: March 2016	168 contributions
Waste of time and energy	You will simply go over to the top of the rice terrace and simply see the green terraces (or I would say just green grasses). No explanation of how rice / pady is prepared. No demo of any manufacturing activity. Infact the hot weather will further frustrate you	Date of experience: October 2017	74 contributions
Diminishing green	You must visit one of the most beautiful place in Bali offering lush green rice terrace and gentle breeze. Jatiluwih...means truly beautiful.	Date of experience: October 2018	244 contributions
everything is just GREEN	You can trek depending to your ability and we only 45 minutes trekking, there are a lot scarecrows in the field to chase away birds. It really is amazing from a scenic as well as historic way of farming point of views.	Date of experience: December 2016	4 contributions
rice terraces for you eyes only	yes, this ia a place to go, and see. You can walk across the rice fields for one hour or more, and then then you must enjoy the panoramic views. Entrey fee: 20.000 IDR	Date of experience: February 2016	1,514 contributions
Wow	Wow, just simply wow...We have visited a few rice fields in Bali but none of them can compare to the fields in this area.. Definitely happy we decided to visit this place	Date of experience: September 2015	8 contributions

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

This place so Geen even in rainy day!	Wow i was surprised how green it can be even we came here in rainy day. Also was interesting to see traditional farming and subak system. So amazing they did it just by hands. Really hard work. Must-visit place!	Date of experience: May 2019	8 contributions

D. Phase -4. Cleansing of Collected Data

Data cleansing is performed to prepare data for actual processing. For cleansing of collected data Natural Language Toolkit (NLTK) has been used. The various tasks related to the cleansing of data involved: removal of extra spaces, removal of punctuations, case normalization, removal of stop-words, tokenization and lemmatization. A brief overview of all performed tasks is given below. Fig.2 shows a raw string extracted from the identified data source which is submitted to NLTK for cleansing and further processing.

```
text = "... spent perusing the exhibitions. Nice     helpful staff,  felt they cared to  keep us safe and were pleased to see us.Thanks.
```

Fig 2. A raw string.

- Natural Language Toolkit (NLTK)

The reviews posted by users are mostly in the English language. The Natural Language Toolkit (NLTK) is a python-based string processing library that takes raw strings as input and removes spaces, punctuations, stop words etc. resulting in a clean string that is more suitable for processing and analysis. Fig.3 shows the output string after the removal of extra spaces from the input string. Fig. 4 shows the output string after the removal of punctuations from the input string.Fig.5 shows the output string after case normalization from the input string.

```
#removing extra spaces

" ".join(text.split())
```

```
'... spent perusing the exhibitions. Nice helpful staff, felt they cared to keep us safe and were pleased to see us.Thanks.
```

Fig 3. String after removing extra spaces.

```
#remove punctuations

from nltk.tokenize import RegexpTokenizer

tokenizer = RegexpTokenizer(r'\w+')
" ".join(tokenizer.tokenize(text))
```

```
'spent perusing the exhibitions Nice helpful staff felt they cared to keep us safe and were pleased to see us Thanks'
```

Fig 4. String after removing extra spaces and punctuations.

```
#case normalization

text.lower()
```

```
' spent perusing the exhibitions nice helpful staff felt they cared to keep us safe and were pleased to see us thanks '
```

Fig 5. String after case normalization.

- **Tokenization:** When pre-processing of reviews is performed, textual data is divided into small tokens and to create a list thereof. Fig.6 shows the word tokenization.

```
from nltk.tokenize import word_tokenize
text = "God is Great! I won a lottery."
print(word_tokenize(text))

Output: ['God', 'is', 'Great', '!', 'I', 'won', 'a', 'lottery', '.']
```

Fig 6. Word Tokenization.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

Fig.7 shows the snapshot of the process of removing the stop words and results

```
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

nltk.download('stopwords')
nltk.download('punkt')

stop_words = set(stopwords.words('english'))

word_tokens = word_tokenize(text)

filtered_sentence = [w for w in word_tokens if not w in stop_words]

filtered_sentence = []

for w in word_tokens:
    if w not in stop_words:
        filtered_sentence.append(w)

print(word_tokens)
print(filtered_sentence)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
['...', 'spent', 'perusing', 'the', 'exhibitions', '.', 'Nice', 'helpful', 'staff', ',', 'felt', 'they', 'cared', 'to', 'keep', 'us', 'safe', 'and',
['...', 'spent', 'perusing', 'exhibitions', '.', 'Nice', 'helpful', 'staff', ',', 'felt', 'cared', 'keep', 'us', 'safe', 'pleased', 'see', 'us.Thank
```

Fig.7 Stop-Word removal script and result.

Lemmatization is performed to find the root word from the meaning or context of the word. The determined root word is called lemma.

Fig. 8 shows the Lemmatization process

```
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

[lemmatizer.lemmatize(word) for word in text.split()]

print("corpora :", lemmatizer.lemmatize("corpora"))
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
corpora : corpus
```

Fig 8. Lemmatization process

E. Phase -5ANALYSES OF DATA

To discover insights that are relevant to objectives stated in phase 1, the collected and pre-processed data needs to be analysed efficiently. The huge amount of collected data is stored in a large number of files and it is impossible to open such large number of files directly onto a system. Feeding of such huge data causes number of problems including problem of processing overhead that takes quite long time. Sometimes it may not be even possible to read, write and process successfully. To cater to the issues a hybrid data processing engine Apache Spark is used. The Apache Spark framework is capable of quickly processing very large data sets, and also distributes data processing task across multiple computers and (or) processors. Fig. 9 shows the overview of the processing of Apache Spark.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

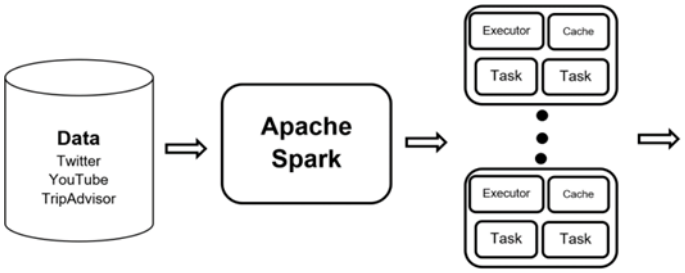

Fig 9. Apache Spark Processing

Apache Sark is then incorporated with Bidirectional Encoder Representations Transformers (BERT) to perform real-time sentiment analysis. BERT Model was pre-trained using IMDB reviews dataset, which is collected and prepared by Andrew L. Maas[40][41]from the popular movie rating service. The model is then fine-tuned as per the set factors discussed in introduction section, on actual dataset. The model is implemented using Keras in TensorFlow. The base variant of BERT sentiment analysis model architecture has 12-layers. Each layer has 768-hidden unit, 12-heads multi-headed attention layer, and can have total of 110M parameters. The dense output layer is stacked with tanh activation function on top of it.

For understanding the context of incoming reviews, the model is pre-trained by token, segment and positional embedding. In Token Embeddings, a CLS is added to the start of the first sentence of a review and SEP is added at the end of the sentence. In segment embedding, a sentence number is assigned to each sentence. Each word in a sentence is assigned a number relevant to its position by positional embeddings.

The pre-trained model divides the task into two sub-tasks: Mask Level Modelling (MLM) and Next Sentence Prediction (NSP). In Mask Level Modelling (MLM) training data is fed with a few tokens being replaced with masked tokens to let the model predict the original word using language and context understanding. Next Sentence Prediction (NSP) is done to know whether the succeeding sentence follows first or not[42]. The other task is to fine-tune the system to make the model learn to solve our problem by feeding actual data.

The outputs from the pre-training phase are class label and a list of word vectors. On this output a SoftMax activation function is applied which contains same number of neurons as in Token library i.e. 30000 in wordpiece library [43] used here to convert these word vectors into distributions. The SoftMax function helps in calculating probability of each class the input belongs.

The actual labels are obtained by a binary classification of each word in the review using one hot encoding as shown in Table 3:

Table 3 One hot encoding of an incoming review.

S.No.	spent	pursing	exhibition	nice	helpful	staff	felt	Cared	keep	us	safe	pleased	see	us	thanks
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

As shown in Table 3, each incoming review is encoded to find as much as possible classes from the data to get the most accurate results. Furthermore, this also increases dimensionality of data to multiple levels and BERT is quite fit for such data.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

The obtained distributions are then compared using cross entropy loss function that focuses on increasing accuracy and thus performs better predictions. The formula for calculating cross entropy loss is calculated by using the formula:

$$loss = -\sum_{i=1}^{m} y_i * \log \hat{y}_i$$

Where, $\hat{y}_i$is the i^{th} value in the model output and, $y_i$is the corresponding target value and m is the number of values in the output.

The proposed algorithm for calculating sentiment score for incoming reviews is shown below.

Algorithm 1: Analysis Algorithm

Input:

HotelsFamily→ Hotel review files with family filter for one destination,
RestaurantsKids→ Restaurant review files with Kids filter for one destination,
HotelsGeneral→ Hotel review files with General Filter for one destination,
RestaurantsVeg→ Restaurant's review files with veg filter for one destination,
Flights → Flight data files,
Destinations → Destination review files,
Twitter → Twitter Data,
Youtube→YouTube Data.
Parameter→ϒ, α, δ

Result:

scores
Initialize: ϒ = 0:5,α = 0:5,δ = 0:5,
Hotel_family_score = 0,
Restaurant_veg_score = 0,
Hotel_general_score = 0,
Restaurant_kids_score = 0,
Destination_score = 0,
Flight_score = 0, twitter_score = 0,
Youtube_score = 0
Repeatforall files;
***for** hf in HotelsFamily do*
    score = sentiment analysis(hf);
    score = sum(score)/len(hf);
    hotel_family_score+ = score;
end
hotel family score/ = len(HotelsFamily);
***for** f in Flights do*
    score = frequency_analysis(f);
    flight score+ = score;
end
flight score/ = len(Flights) ;
veg_food_score = restaurant_veg_score ;
*facilities_score = (δ * hotel_general_score +ϒ *restaurant_veg_score)/ (ϒ + δ);*
*comfort_score = (δ * hotel_general_score+ ϒ * flight_score) / (ϒ + δ);*
*destination_score = (δ * destination_score+ ϒ * twitterscore+ α *youtubescore)/ (α + ϒ + δ);*
*family_score = (δ * restaurant_kids_score + ϒ * hotel_family_score) / (ϒ + δ);*
return
scores

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

For optimization Adam Optimizer [44] is used which is abbreviated as adaptive moment estimation. As Adam is capable of working with huge amount of data categorized among large number of parameters makes it is easy to implement, computationally efficient and requires less memory space[45], and thus it is perfect choice to be chosen for optimization of our problem.

Sentiment scores are obtained by using positive and negative words and then normalizing these scores. This sentiment score is finally analysed using a suitable index as discussed in experimental setup and execution section.

.

VII. EXPERIMENTAL SETUP AND EXECUTION

The proposed framework has been applied on the data, related to the tourism, available on various social media sites in the form of text reviews. The details of the experimental setup are shown in Table 4. It includes data sources, destinations covered, number of extracted reviews (consolidated and destination wise) and various attributes that are under exploration.

Table 4. Experimental Setup Details.

Date Sources	TripAdvisor.com, Twitter, YouTube
Data Type	Text in Natural Language
Places Under Consideration	Bali, Bangkok, Dubai, London, Mauritius, New York, Paris, Singapore, Sri Lanka, Switzerland. (Total 10 Places)
Data Extraction Period	October –2021
Total No. of Reviews Collected	96,561 Reviews.
Place-wise no of Reviews collected	
Bali	9,574 Reviews
Bangkok	10,000 Reviews
Dubai	9,998 Reviews
London	10,000 Reviews
Mauritius	8,351 Reviews
New York	9,924 Reviews
Paris	9,982 Reviews
Singapore	10,000 Reviews
Sri Lanka	9,139 Reviews
Switzerland	9,596 Reviews
Attributes over which reviews were analysed	
Comfort, Destination, Facility, Family Friendliness, Veg Food.	

The extracted reviews were passed to BERT model. The probabilistic value is generated by applying the embeddings on the reviews. The generated probabilistic values are evaluated to get the sentiment score of each incoming review. After this processing, a consolidated file was generated for each attribute of every destination leading to the creation of 10 X 5 = 50 files. A snapshot of the output file is as shown in Table 5.

Table 5. A snapshot of Output Data.

title	text	date_of_stay	helpful_votes	sentiment	score
Highlight of my trip!	You will simply go over to the top of the rice terrace and simply see the green terraces (or I would say just green grasses). No explanation of how rice / pady is prepared. No demo of any manufacturing activity. Infact the hot weather will further frustrate you	Date of experience: March 2016	168 contributions	POSITIVE	0.999430358
Waste of time and energy	You will simply go over to the top of the rice terrace and simply see the green terraces (or I would say just green grasses). No explanation of how rice / pady is prepared. No demo of any manufacturing activity. Infact the hot weather will further frustrate you	Date of experience: October 2017	74 contributions	NEGATIVE	0.999981761
Diminishing green	You must visit one of the most beautiful place in Bali offering lush green rice terrace and gentle breeze. Jatiluwih...means truly beautiful.	Date of experience: October	244 contributions	POSITIVE	0.997407138

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

		2018			
everything is just GREEN	You can trek depending to your ability and we only 45 minutes trekking, there are a lot scarecrows in the field to chase away birds. It really is amazing from a scenic as well as historic way of farming point of views.	Date of experience: December 2016	4 contributions	POSITIVE	0.98627311
rice terraces for you eyes only	yes, this ia a place to go, and see. You can walk across the rice fields for one hour or more, and then then you must enjoy the panoramic views. Entrey fee: 20.000 IDR	Date of experience: February 2016	1,514 contributions	POSITIVE	0.999802053
Wow	Wow, just simply wow...We have visited a few rice fields in Bali but none of them can compare to the fields in this area.. Definitely happy we decided to visit this place	Date of experience: September 2015	8 contributions	POSITIVE	0.999715984
This place so Geen even in rainy day!	Wow i was surprised how green it can be even we came here in rainy day. Also was interesting to see traditional farming and subak system. So amazing they did it just by hands. Really hard work. Must-visit place!	Date of experience: May 2019	8 contributions	POSITIVE	0.999361455

After obtaining output files having the positive and negative sentiment scores, a single value based upon both sentiment score and view count need to be generated. The Relative Strength Index (RSI) [46] is determined as a net oscillation of the positive/ negative momentum. The formula for RSI is as follow:

$$RSI = 100 - \left(\frac{100}{1 + (Average\ Positive\ sentiment)/(Average\ Negetive\ sentiment)}\right)$$

VIII. RESULTS

The sentiment scores computed to determine a single value preference index for each attribute/place have been provided to indicate the tourist experience. Fig.10 shows relative sentiment index for comfort attribute across various destinations. It indicates that Bali is the most comfortable place, whereas Singapore is least comfortable place among all considered places. Fig.11 shows relative sentiment index for sightseeing attribute. It indicates that Sri-Lanka is the least preferred place as a destination to travel for the purpose of sightseeing while Switzerland is the most favourite destination.

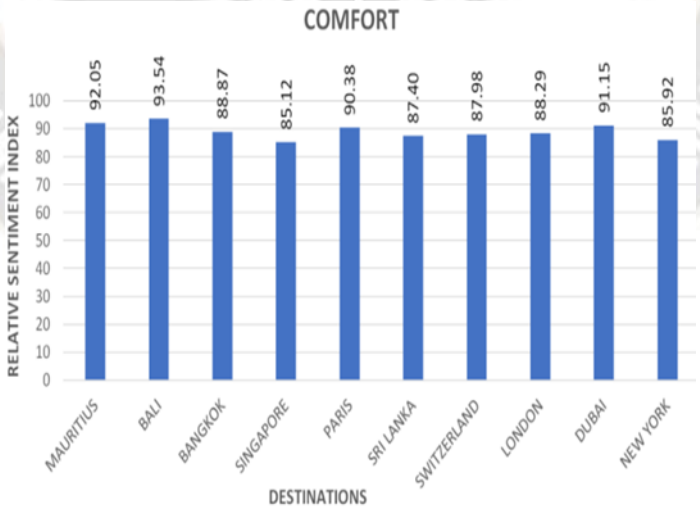

Fig.10 Relative Sentiment index for comfort across places.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

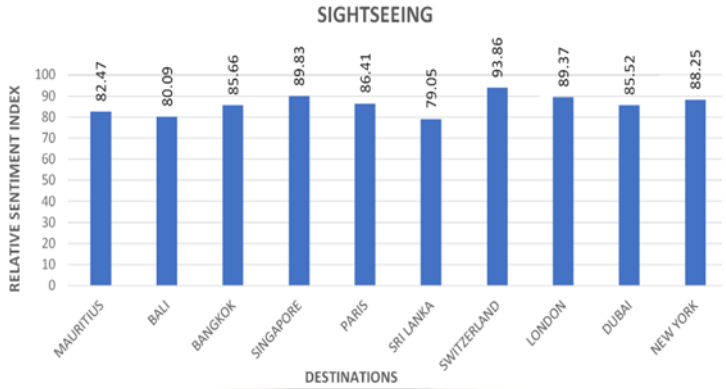

Fig. 11 Relative Sentiment index for sightseeing across places

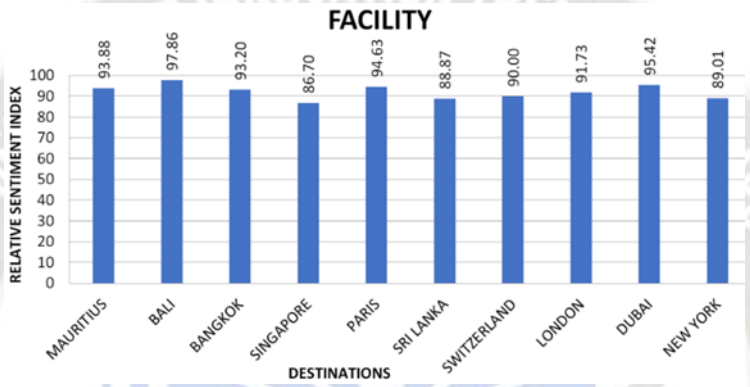

Fig.12 Relative Sentiment index for facilities across places.

Fig. 12 shows relative sentiment index of facility attribute. It shows that Bali provides the maximum facilities while Singapore has the least facilities among all considered places. Fig. 13 represents the relative sentiment index for Family friendliness attribute indicating how favourable a place is from the viewpoint of family enjoyment. New York being the least favourable while Bali being the most favourable for families to visit among the places taken into consideration.

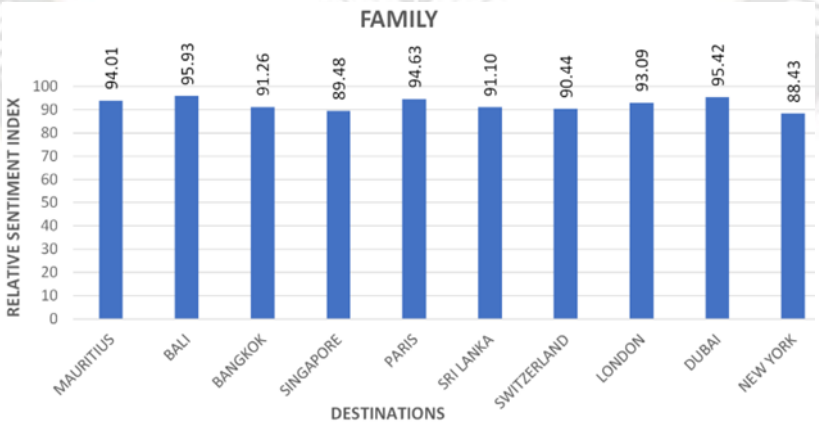

Fig.13 Relative Sentiment index for Family friendliness across places.

Fig. 14 represents the relative sentiment index for Vegetarian food availability attribute.The graph suggested that most of the places offer good availability of veg food. Paris is the most favorable in the group while Singapore is the least favorable for vegetarians

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
Article Received: 14 December 2022 Revised: 22 January 2023 Accepted: 29 January 2023

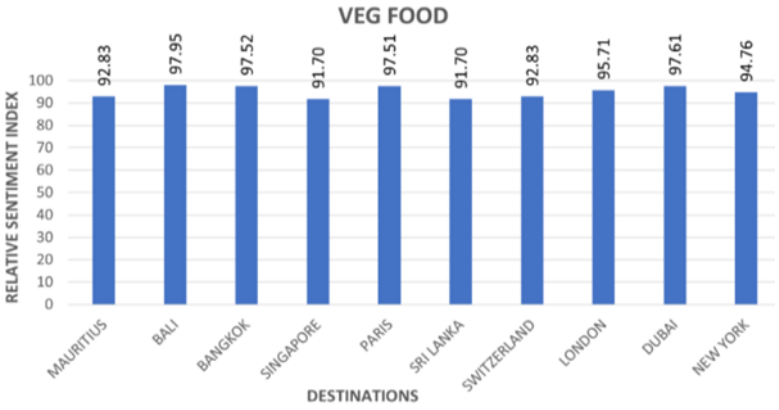

Fig.14 Relative Sentiment index for Veg food availability across places.

Fig. 15 represents a consolidated graph that compares various places for different attribute values. This will help the user in making an appropriate choice for a tourist destination on the basis of a large number of reviews expressed by people .

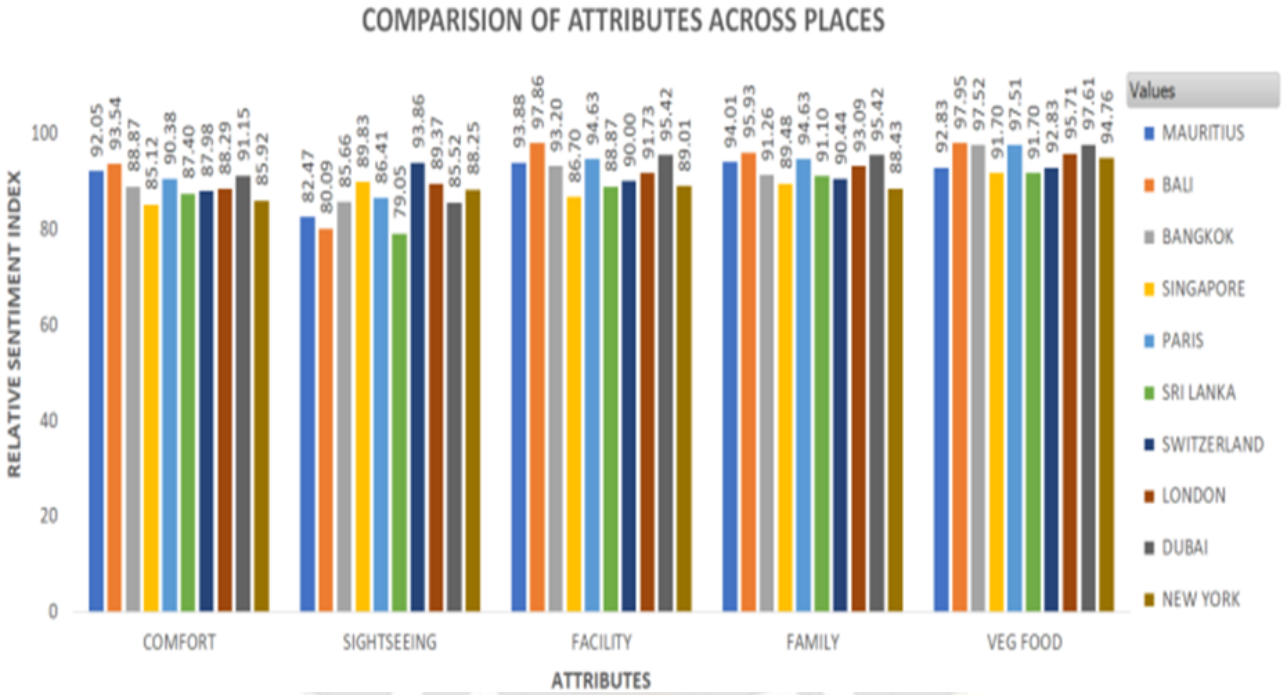

Fig.15 Consolidated graph for all places and attributes.

After exhibiting the above-shown information in the form of a graph, the proposed model was extended to display the above information on the World map in an interactive manner.

IX. VISUALIZATION OF RESULTS

Fig 16 provides visual presentation of results on world map implemented using JavaScript and hover effect. A zoom into place

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

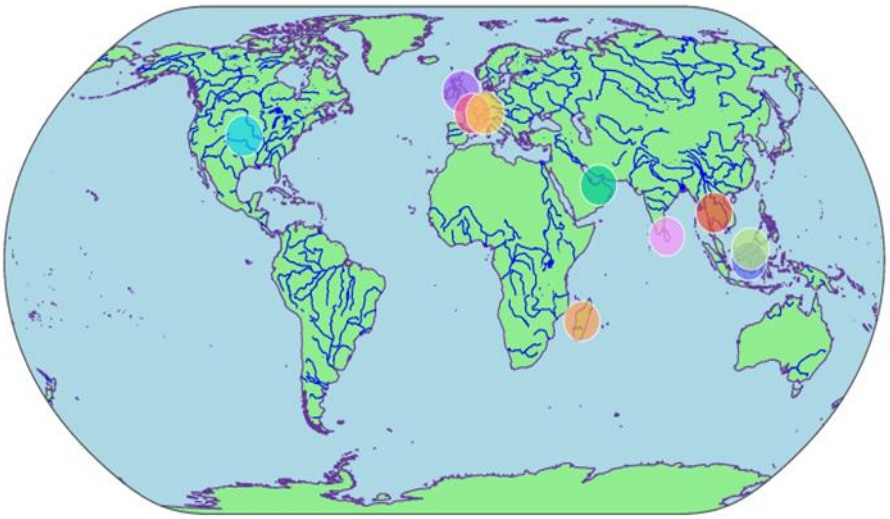

Fig.16 Display of final sentiment score on world map

facility has also been provided to zoom into any place in the world map. As mouse curser is placed over the coloured dot relative sentiment score of all the attributes that destination is displayed into same coloured box as of dot and displays a summary of that place. Fig. 17 represents zoomed view of relative sentiment score for Sri Lanka.

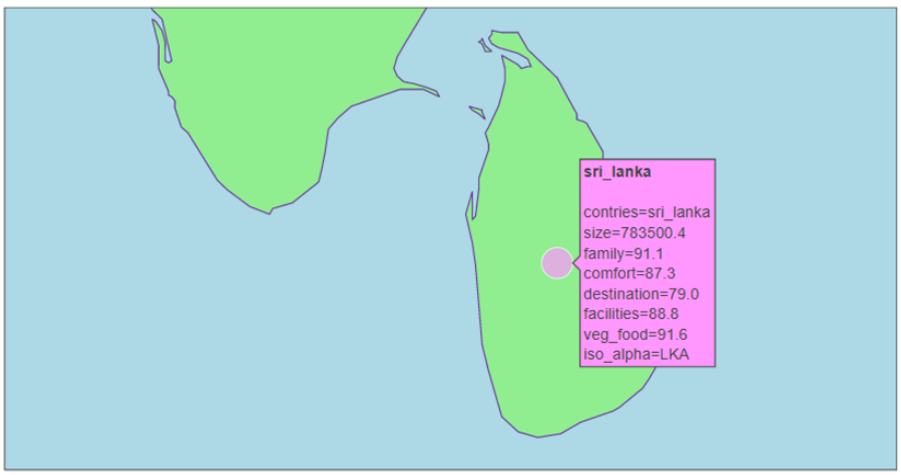

Fig.17 Relative sentiment score of Sri Lanka on World map in zoomed view.

X. CONCLUSION AND FUTURE SCOPE

The work carried out in this paper implements a Big Data Analytics based framework for consolidating and interpreting the tourist recommendations in the form of numerical score after processing the textual data for large number of reviews which is manually not feasible. The work involves the extraction of reviews from various social media sources, evaluation of each review for its sentiment type and extent in the form of sentiment score, combining all the positive and negative sentiment scores to a single value decomposition using RSI. The output of the work has been exhibited on the world map in the interactive manner which can be seen through the movement of the mouse. The work has been carried out for ten most popular destinations and can be extended for all the major destinations across the globe. The number of attributes can also be extended.

REFERENCES

[1] https://www.bbc.co.uk/bitesize/guides/zqk7hyc/revision/1
[2] https://www.statista.com/statistics/207009/number-of-outbound-visits-of-indian-nationals-from-india-since-2000/
[3] https://timesofindia.indiatimes.com/business/india-business/spends-on-foreign-trips-up-253x-in-5-yrs/articleshow/65815814.cms

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

[4] https://timesofindia.indiatimes.com/business/india-business/desis-q1-foreign-trip-bill-at-new-high/articleshow/70693721.cms
[5] https://www.makemytrip.com
[6] https://www.tripadvisor.in
[7] https://www.easemytrip.com
[8] https://www.yatra.com/
[9] https://twitter.com/login
[10] https://www.facebook.com
[11] https://www.instagram.com
[12] http://www.walkthroughindia.com/around-the-world/top-15-most-visited-countries-by-indian-travellers/
[13] https://www.selenium.dev/downloads/
[14] https://pypi.org/project/selenium/
[15] https://beautiful-soup-4.readthedocs.io/en/latest/
[16] https://www.crummy.com/software/BeautifulSoup/bs4/doc/#installing-beautiful-soup
[17] https://spark.apache.org/docs/latest/
[18] https://www.nltk.org/
[19] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html
[20] Ana Reyes-Menendez , Jose Ramon Saura, And Juan Gabriel Martinez-Navalon, June, 2019, The impact of e-wom on hotels management Reputation: Exploring TripAdvisor Review Credibility with the ELM model, Translation And Content Mining, IEEE Access, Vol 7, 2019.
[21] S. M. Al-Ghuribi, S. A. Mohd Noah and S. Tiun, "Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews", IEEE Access, vol. 8, pp. 218592-218613, 2020., doi: 10.1109/ACCESS.2020.3042312.
[22] K. S. Ntalianis, A. Kener and J. Otterbacher, "Feelings' Rating and Detection of Similar Locations, Based on Volunteered Crowdsensing and Crowdsourcing," in IEEE Access, vol. 7, pp. 90215-90229, 2019, doi: 10.1109/ACCESS.2019.2926812.
[23] Benlahbib and E. H. Nfaoui, "Aggregating Customer Review Attributes for Online Reputation Generation," in IEEE Access, vol. 8, pp. 96550-96564, 2020, doi: 10.1109/ACCESS.2020.2996805.
[24] S. M. Al-Ghuribi and S. A. Mohd Noah, "Multi-Criteria Review-Based Recommender System–The State of the Art" in IEEE Access, vol. 7, pp. 169446-169468, 2019, doi: 10.1109/ACCESS.2019.2954861.
[25] İ. Topal and M. K. Uçar, "Hybrid Artificial Intelligence Based Automatic Determination of Travel Preferences of Chinese Tourists," in IEEE Access, vol. 7, pp. 162530-162548, 2019, doi: 10.1109/ACCESS.2019.2947712.
[26] K. A. Fararni, F. Nafis, B. Aghoutane, A. Yahyaouy, J. Riffi and A. Sabri, "Hybrid recommender system for tourism based on big data and AI: A conceptual framework," in Big Data Mining and Analytics, vol. 4, no. 1, pp. 47-55, March 2021, doi: 10.26599/BDMA.2020.9020015.
[27] G. Adomavicius and A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.
[28] G. Sun, R. Liang, H. Qu and Y. Wu, "Embedding Spatio-Temporal Information into Maps by Route-Zooming," in IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 5, pp. 1506-1519, 1 May 2017, doi: 10.1109/TVCG.2016.2535234.
[29] J. He and H. Hu, "MF-BERT: Multimodal Fusion in Pre-trained BERT for Sentiment Analysis," in IEEE Signal Processing Letters, doi: 10.1109/LSP.2021.3139856.
[30] H. C. M. Senefonte, M. R. Delgado, R. Lüders and T. H. Silva, "PredicTour: Predicting Mobility Patterns of Tourists Based on Social Media User's Profiles," in IEEE Access, vol. 10, pp. 9257-9270, 2022, doi: 10.1109/ACCESS.2022.3143503.
[31] P. Nitu, J. Coelho and P. Madiraju, "Improvising personalized travel recommendation system with recency effects," in Big Data Mining and Analytics, vol. 4, no. 3, pp. 139-154, Sept. 2021, doi: 10.26599/BDMA.2020.9020026.
[32] Y. Wang, M. Wang, and W. Xu, ``A sentiment-enhanced hybrid recommender system for movie recommendation: A big data analytics framework," Wireless Commun. Mobile Comput., vol. 2018, Mar. 2018, Art. no. 8263704.
[33] S. Aciar, D. Zhang, S. Simoff, and J. Debenham, ``Informed recommender: Basing recommendations on consumer product reviews," IEEE Intell. Syst., vol. 22, no. 3, pp. 39_47, May/Jun. 2007.
[34] L. Sun, J. Guo, and Y. Zhu, ``Applying uncertainty theory into the restaurant recommender system based on sentiment analysis of online Chinese reviews," World Wide Web, vol. 22, no. 1, pp. 83_100, 2018.
[35] M. Siering, A. V. Deokar, and C. Janze, ``Disentangling consumer recommendations: Explaining and predicting airline recommendations based on online reviews," Decis. Support Syst., vol. 107, pp. 52_63, Mar. 2018.
[36] R. Dong, M. P. O'Mahony, M. Schaal, K. McCarthy, and B. Smyth, ``Sentimental product recommendation," in Proc. 7th ACM Conf. Recommender Syst., 2013, pp. 411_414.
[37] G. Chen and L. Chen, ``Recommendation based on contextual opinions," in Proc. Int. Conf. Modelling, Adaptation, Pers., 2014, pp. 61_73.
[38] X. Zheng, Y. Luo, L. Sun, J. Zhang, and F. Chen, ``A tourism destination recommender system using users' sentiment and temporal dynamics," J. Intell. Inf. Syst., vol. 51, no. 3, pp. 557_578, 2018.
[39] H. S. Indriany, A. N. Hidayanto, L. J. Wantania, B. Santoso, W. U. Putri and W. Pinuri, "Data Quality Management Maturity: Case Study National Narcotics Board," 2021 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), 2021, pp. 206-212, doi: 10.1109/COMNETSAT53002.2021.9530824.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 3s
DOI: https://doi.org/10.17762/ijritcc.v11i3s.6176
***Article Received**: 14 December 2022 **Revised**: 22 January 2023 **Accepted**: 29 January 2023*

[40] https://ai.stanford.edu/~amaas/data/sentiment/
[41] https://www.kaggle.com/rudrasing/andrew-mass-dataset
[42] B. Wang and C. J. Kuo, "SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2146-2157, 2020, doi: 10.1109/TASLP.2020.3008390.
[43] https://cran.r-project.org/web/packages/wordpiece/vignettes/basic_usage.html.
[44] Z. Zhang, "Improved Adam Optimizer for Deep Neural Networks," 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), 2018, pp. 1-2, doi: 10.1109/IWQoS.2018.8624183.
[45] Z. Zhang, "Improved Adam Optimizer for Deep Neural Networks," 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), 2018, pp. 1-2, doi: 10.1109/IWQoS.2018.8624183.
[46] H. Shehata and T. Khattab, "Energy Detection Spectrum Sensing in Full-Duplex Cognitive Radio: The Practical Case of Rician RSI," in IEEE Transactions on Communications, vol. 67, no. 9, pp. 6544-6555, Sept. 2019, doi: 10.1109/TCOMM.2019.2916069.