National College of Ireland

Behaviour Patterns from Social Media Images - An ideal way for Tourism Marketing?

Research in Computing
MSc Data Analytics

Aniket Bhawkar
Student ID: x17170885

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Aniket Bhawkar
Student ID:	x17170885
Programme:	MSc Data Analytics
Year:	2018-19
Module:	Research in Computing
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	12/08/2019
Project Title:	Behaviour Patterns from Social Media Images - An ideal way for Tourism Marketing?
Word Count:	6847
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	3rd August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

Behaviour Patterns from Social Media Images - An ideal way for Tourism Marketing?

Aniket Bhawkar
x17170885

Abstract

Digital mediums are being used far more often throughout the globe for sharing images. Channels like Facebook, Instagram, Twitter and Flickr are regularly used by people to publish their travel images. These pictures can elaborate a pattern in the travel preferences and choices of the users. The purpose of this research is discovering the behavioural patterns of users towards travel from the uploaded social media images with the use of computer vision technology and using this information for tourism marketing. Computer vision tools like Amazon Rekognition, Google Cloud Vision and Cloud Sight enable feature extraction from the images. These features include highlights, visual features, content features and labels. Owing to the outcome, algorithms like K-Nearest Neighbour(KNN) and K-Means could be useful in classifying the behavioural patterns. Recurrent Neural Network - Long Short-Term Memory(LSTM) could be effective in evaluating the change of user preferences over a period of time and predicting the next desirable choice. The results can be compared geographically, with which the system can then generate recommendations more effectively. Content-based and Collaborative Filtering Recommendation technique would be used in the proposed model. The information received from the model could be then used for marketing tourism packages.

Keywords*: behaviour patterns, recommendation system, tourism, marketing*

1 Introduction

Advancement of technology has enabled humans to commute from one corner of the world to the other in a short duration. Several options of transportation are available to facilitate the feasibility of choices. The world wide web provides numerous go-to options and alternatives which could be considered depending upon personal preferences. This causes a roadblock to the choosing capability of a potential tourist in identifying their perfect vacation destination. Online booking websites have influenced the out-performance of the traditional ticket way of travel package booking method, which involved the customer visiting the travel agent and selecting the best package as per their budget and preferences. In online booking, people lay down their expectations while viewing the images provided over the website Narangajavana et al. (2017). Hence, a potential tourist may have queries like where, how and when to travel? Nowadays, online portals like Trip-Advisor, Expedia and Airbnb grill down information in solving these quests to narrow the choices of travelling.

Pictures can convey a million words. Hence social media platforms have made the feature to upload images as their integral part. People make us of these image-based platform to its fullest and express their emotions through the images which they upload Wang and Li (2015). The monthly active Facebook users worldwide as of fourth quarter 2018 was over 2,320 million [1], whereas that of Twitter was 326 million [2] and those on Instagram were 1,000 million [3]. Image based platform like Instagram and Flickr have clinched popularity over the previous decade which is used to share travel pictures [4].

Its the human tendency to get pulled in to things and spots their family or companions have visited Kaosiri et al. (2019). Succeedingly, a potential tourist may choose to explore a different location. Hence, a travel agent faces the dilemma of which package to recommend therefore arising the need of a model which can predict the keen interest of a potential tourist. Owing to this difficulty, deeper insights about the behavioural patterns of that person could be fetched from social media. This content is known as User Generated Content (UGC). The travel images posted over these mediums can provide the behaviour, personality and their preferences Ferwerda and Tkalcic (2018). Furthermore, the obtained information could be utilised for recommending personalised travel packages.

With the growing affection of tourism, cross-cultural interactions between different cultural groups are increasing. It is observed that cultural background can affect the choices made by a potential customer Filimonau and Perez (2018). International vacation is often being pursed by a sparse population who are wealthy enough to afford it Ahn and McKercher (2015). Therefore, there is a need to analyse the effect of cultural background in making travel choices. Hence, travel agencies shortfall to perceive the hostile effect of behavioural patterns thus leading to low package sales.

Computer vision tools like Google Cloud Vision, Cloud Sight and Amazon Rekognition are sophisticated image processing and labelling tools which can recognize the content and provide labels Mulfari et al. (2016). They are capable to classify images into groups and provide appropriate labels, extract textual information, identify objects, faces and facial expressions Bosch et al. (2018). These technologies facilitate with a REST API with which the service could be accessed. Every technology has few limitations; extrapolating percent sentiments is not available for public and is in the beta state which is unavailable in the developers API; Moreover, Google Cloud Vision was proved to be prone to blur and noise Hosseini et al. (2017).

In order to maximize conversions, it is highly reasonable for the travel agents to identify the keen interest of the tourist for providing custom personalised packages. Hence, to attain this target and promote packages more efficiently; there is a need for a model to predict the behaviour patterns Shiranthika et al. (2018). The proposed research in intended to identify the keen interests of a potential client using their social media pictures with the help of computer vision labelling techniques. Data mining of these labels can provide a base for the travel agents to market the packages in more personalised manner. Density-based spatial clustering algorithm(DBSCAN) can identify the popularity of a location and the prefered geo-location of a user; this information can be fetched from the geo-tagged pictures Ester et al. (1996). Machine learning algorithms like KNN and K-means would be used over the encoded labels; the results would then predict the change of behavioural patterns over a period of time using LSTM algorithm. The machine learning

[1]https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

[2]https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[3]https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/

[4]https://www.statista.com/statistics/625240/most-instagrammed-tourist-attractions/

model would exclusively involve the factors like Popularity, Filtering techniques (Collaborative Filtering and Content-Based Filtering), Geo-Locations and External Factors (Seasonal).

From the results, the travel industry could change their marketing campaigns as per the needs to enhance consumer engagement which will inturn increase sales of packages. This research paper comprises of Literature Review (section 2), Proposed Methodology (section 3), Experimentation (section 4), Proposed Evaluation (section 6), Discussion (section 7), Ethics (section 8) and Conclusion (section 9).

2 Literature Review

Recommendation of the packages are usually carried out on the brief given by a probable tourist while considering the traditional method of ticket booking. Be that as it may, with the development of online reservation systems, travel booking are majorly done over the internet in the past decade which has brought about an outstanding breakdown in the conventional ticket reservation process. Therefore, it's necessary to highlight the packages which would be preferred by the potential client, in order to enhance the customer conversion. The report consist of literature review which is sub-sectioned into 2.1 Behaviour Pattern Identification, 2.2 Travel Spot Recommendation Systems and 2.3 Factors Impacting Tourism.

2.1 Behaviour Pattern Identification

Social media platforms have provided the feature to upload unstructured data which is being used widely by people. This data is usually in the form of images, videos, text, notes, etc and therefore increases the need to study and mine this big data for deducing the behaviour preferences and patterns of the users Miah et al. (2017). The tourism business is in boom which increases the need to improve preferences analysis to order to enhance business decisions. Miah et al. (2017) developed a model which was built specifically for Melbourne, Australia in order to understand unstructured big data over social channels to improve decision making. Textual data in the form of caption and comments could be analysed, geolocation detection, image analysis and time-series analysis are some of the major areas covered in this research. On the basis of these tests, the keen interest of candidates, geolocation clustering, trend and seasonal modelling could be discovered. On these bases, the demand of tourism and their trends could be optimized more efficiently, traffic and transportation, hospitality and accommodation services could also be well organised. However, the article could be improved by evaluating the influence of cultural differences.

Social media platforms, nowadays allow users to upload textual and images in order to express their feelings. Instagram and Flickr being leading image-based platforms are more popular for travel, fashion, architectural, etc content. Google Cloud Vision API is used to extract visual and content features from the uploaded data Ferwerda and Tkalcic (2018). The authors trained the model using M5' rule, random forest classifier and radial basis function network with respect to predicting the personality of an Instagram user based on the images which they have uploaded. As per Ferwerda and Tkalcic (2018), the experiment justified visual and content features can standalone display significant results. This experiment elaborates the use of computer vision tools for extracting features from

images and then after the content could be used for machine learning techniques in personality detection.

User-Generated Content(UGC) is generated tremendously over social media channels like Twitter. It was a challenge in understanding the behaviour patterns over Twitter which is dynamic in nature and contains more content in the form of trends. Hasnat and Hasan (2018) extracted geo-tagged tweets of users with respect to tourism in understanding their travel preferences. This survey was based in Florida, USA and extracted geo-locational travel data. A pattern of tourist areas visited by an individual was shaped and this information was passed to a Conditional Random Field (CRF) to foresee the succeeding possible destination. Based on the information analysed, the model predicted the accuracy of 64.9% on an average while predicting the anticipated travel location. The conventional travel locations which were considered included spots like Arena, University, Shopping Centre, Convention Centre, Lake / River, Stadium, Cemetery, Zoo and Hospitals. These patterns produce an understanding of the travel preferences of a social media user and the choices they possess. This research provides a low-cost method of analysing the behaviour patterns from the UGC over social media and attempts in predicting their next travel destination.

E-commerce platforms, for improving their sales needs to identify the behaviours of the potential customers. For achieving the results, GERU et al. (2018) collected 900 Instagram images entitled under the hashtag #thegoodlife. Computer vision tools were used to label these images. K-Means algorithm was then used on these labels to identify the behaviour patterns. Clusters are formed using the K-Means algorithm, due to which dependency between labels and the way they are inter-related with each other could be distinguished. The labels can then be used for keyword-based targeting, most of the social media platforms enable this feature. GERU et al. (2018) aimed in understanding the patterns from Instagram images could be used to e-Commerce marketing. This model could be more statistically evolved by performing a Chi-Square test, in which the tags mentioned in the caption and those received from the computer vision tools could be compared. The received results were then compared with the existing state-of-the-art marketing techniques. This model could be improved by including geo-location, comments, sentiment score, regional impacts, presence of a brand, etc.

As seen earlier, the labels fetched from the computer vision tools are vital in training a machine learning model. Hosseini et al. (2017) noticed that computer vision tools are not accurate where the image contains noise. On further analysis, when testing the original image prior to the noisy image, the tool was able to produce accurate results. For this experiment, Hosseini et al. (2017) made use of Google Cloud Vision and concluded that it was adaptive to filtering and a quick learner. They also added that there's no need for a major upgrade in the algorithm. These tools were capable to produce highly accurate results. In a similar experiment, Bosch et al. (2018) analysed 1818 images over the Google Cloud Vision and it was able to provide the labels within minutes in a cost-effective way while a normal human could need more than 35 hours to generate the same results.

The above works of literature signify the importance of textual data mining, geo-locations, image feature extraction and time-series analysis in the process of identifying user behaviours and seasonal trends. This information can be used for tourism marketing. The classification of the data could be done using various machine learning algorithms. KNN, K-Means, M5' rule, Random Forest Classifier is useful in this scenario. A cloud-based hosting would serve the storage requirement. Cloud storage systems are reliable, scalable, easily accessible and dependable. Various data analytics tools and models are

available over the cloud networks where the necessary data could be used for analysis Yetis et al. (2016). This environment would be useful in the storage of massive big data received from the computer vision tools, text analysis, image database and weather API.

2.2 Travel Spot Recommendation Systems

The Travel Spot Recommendation Systems are algorithms which use the data provided by the potential client in generating preferences. These systems are usually categorised into the following types:

- Collaborative Filtering - People with similar interests are included in this category.
- Content-Based Filtering - Content uploaded is taken into consideration for providing recommendations.
- Hybrid Filtering - It incorporates both Collaborative and Content-Based Filtering.

Nowadays the majority of the UGC is being produced over Facebook which is in the form of images, comments, likes, posts, status updates, etc. Shiranthika et al. (2018) proposed a research which utilized this information for providing personalised recommendations with respect to content-based and collaborative filtering techniques. Labels were being generated from these images using computer vision tools. Geo-locations, age group, sexuality were some parameters which were fetched along with other supporting parameters like likes and relationship were also fetched using the Graph API. KNN, K-Means and Hierarchical clustering algorithms were used in the implementation of this research. Owing to the experiment, the classification of a user was done. This model could be improvised further by using hybrid filtering recommendation technique. A time-series analysis of the profile would provide insights about the seasonal trends and provide better recommendations.

The content which is uploaded over the social media channels could be geo-tagged. As per Memon et al. (2015), the photos which consist of geo-tagged location play a crucial part in providing travel recommendations. It was also mentioned that other dependent parameters like time, weather, title and tags could also create an impact on the results. The model generated was capable to produce accurate results based on the time and other preferences of the potential client. Content-based filtering technique was performed in this system, which produced results based on the users' previous social media data uploaded. The model predicted travel locations and would also recommend famous places or unalike and finer recommendations. Moreover, the architecture used for developing this model could be reused in the proposed system and could be utilised to satisfy the requirements.

The geo-location received could be utilised to develop an itinerary recommendation system with the help of a pattern mining strategy Cai et al. (2018). The system took into consideration various constraints and generated a customized itinerary as per the user request. It made use of sequential, aspatial semantics and user preferences in order to provide the necessary results. The Region of Interest(RoI) was being used as the data mining algorithm from analysing the RoIs from the course of the movement. The RoIs are then converted into a sequence of places and likewise depending on the day type, city, climatic conditions and timings the recommendations were provided. These parameters were used to increase the accuracy of the model. Random selection and Popularity based selection methods were utilised in this experiment. Top five candidates itinerary

recommendations were considered as a final recommendation. The research lacked in considering the time-series behaviour of user preferences.

As stated by Jiang et al. (2016), images uploaded on social media channels are linked with heterogeneous metadata information which could be utilised in generating customised travel recommendations. The proposed model was developed with the intention of producing a personalised recommendation sequence of travel based on the Point of Interest(PoI). Community and travellers pictures which were uploaded on Flickr were used in developing the model which evaluated 7,387 users from Flickr comprising of 7 million images and over 24 thousand travelogues focusing on 864 travel PoIs within 9 popular cities. Jiang et al. (2016) model provided a notable result depending on the data, however, the system could be enhanced over the dependencies which were the timings of PoI which were related to the user experiences. Also, the model could consider the transportation and hotel availability to magnify user experience.

The history of geo-location details fetched from the social media channels are known as Digital Footprints can be used for determining the user preferences Yu et al. (2016). In other words, this flow is known as Location-Based Social Networks (LBSNs). The model utilised Collaborative filtering technique and LBSNs from the collected data to recommend customised travel packages based on user preferences. The system also considered visiting sequences and POIs were being considered while producing recommendations. The authors developed a user preference modelling algorithm to tackle the problem of user preferences, Location Popularity and Location Dependency for location modelling and also, a Point of Interest Discovery and Travel Route Planner model for enhancing data discovery. Crowd-Sourced digital footprints generated along a course of time were used in recommending travel packages. Furthermore Sun and Lee (2017) suggested that accuracy of the travel route planner could be improved by considering other dependent parameters like time of visit, stopover time, duration, the sequence of visiting locations, etc.

The research held by Ge and Persia (2019) stated that most of the recommendation systems available for focus on only one aspect and the often fail to validate other crucial factors. These factors include 1) Social relations, 2) Evaluating methods, 3) Emerging social media, 4) Credibility and 5) Privacy issue. An ideal recommendation engine which is social media-centric could be developed based on these aspects. These aspects are correlated with each other and could impact the integrity of the model. Ge and Persia (2019) signified the importance and the process in which an ideal social media recommendation system is to be progressed. These factors cannot be neglected as they can have an adverse impact on the recommendation system.

From the above works of literature, the process of development of a travel recommendation system could be perceived. The importance of digital footprints and geotagged images illustrate the PoI of the potential customer and their preferences could be evaluated. Machine Learning algorithms like KNN, K-means clustering could be utilised to get notable results and hence are considered in the proposed research. Metadata information, time-series data, weather, user preferences, geo-location information can also be considered to improve accuracy.

2.3 Factors Impacting Tourism

While making a decision, the UGC can have an effect over the choices made by a potential client. Kaosiri et al. (2019) noticed the impact of tourist satisfaction over their pre-

purchase, during vacation and post-vacation period. Narangajavana et al. (2017) pointed out these effects of pleasure and displeasure. It was also observed by Kaosiri et al. (2019) that there is a direct correlation of the choices made by friends, acquaintances, family members, internet users and travel agents. These relationships were then categorised into strong-tie and weak-tie sources. It was observed that these tie sources do have an impact on the anticipation of a travel destination as they usually enlighten the attractions, culture, must visit places, etc. Kaosiri et al. (2019) also justified the importance of sentiments present over the UGC with respect to the level of anticipation a would-be tourist have and its effect on the travel industry. Moreover, the impact can be improvised by taking into consideration the effect of influencers and trends.

As seen in the previous research, online images can cause an influence on the expectations of a tourist based on the source of the tie. However, similar research was conducted by Lian and Yu (2019) to identify the potential risks and their adverse impacts on the expectations virtue of the online images. It was also necessary to keep in a record, the influence of image source, image quality, and other metadata values which can knock the decision making capability of a potential tourist. Lian and Yu (2019) focused on a short survey which was based on Huangshan Mountain Scenic Spot, which included a total of 240 adult participants under 40 years of age. From this experiment, the authors concluded that the source of images does not have a significant impact, but the direction and quality of the image can influence the decision making capability. Hence if is necessary to make of quality images which developing a travel recommendation system. Social media is being immensely used to increase brand value and tourism of a particular location Huertas and Marine-Roig (2016). Hence it is necessary to make use of more location specific content and quality images.

The cultural background is an important factor which can affect the tourism choices Filimonau and Perez (2018). It's essential for a travel agent to consider the cross-cultural interactions in order to enhance the selection process. Therefore, a travel agent must always study the cultural background of a potential client before recommending them travel packages. Hofstede's cultural dimensions model was utilised to understand the influence of cultural diversities over the travel preferences Filimonau and Perez (2018). The authors conducted a study on the tourist who belongs to either the United Kingdom and Venezuela. From this study, the authors were able to recognize minor similarities, although a notable distinction was present in the pre-holiday and during the holiday period. The authors suggested developing a machine learning model which could distinguish these behaviour based on the factors available.

Based on the results, the preferences of the customer can be identified. The algorithms can then distinguish the user in k distinct groups with a similar set of interest and provide recommendations accordingly. For this project, people can be categorized into Nature Lover, Beach person, Adventurous, Fashion Freak, Explorer and so on [5].

It's necessary to identify how images uploaded over social media channels could be utilised in understanding the travel preferences - behaviour patterns and can these outcomes be utilised for tourism industry marketing?

Social media platforms have provided the feature with which the audience can express their sentiments in the form of pictures and textual data Ferwerda and Tkalcic (2018). This information provides attributes like geo-location, caption, age, gender, profession, number of likes, comments, time, date, day, weather, season and much more which could be useful in developing a machine learning model. These images would be compiled us-

[5]https://www.cntraveller.com/gallery/10-types-that-travel

ing computer vision tools and the extracted labels and meta information. The labels which are extracted are diverse in nature and may differ and hence it is necessary to have them categorized; a data dictionary could be used in categorising this information. Thereafter, machine learning techniques like KNN, K-Means could be used on this information. RNN-LSTM can predict time-series patterns. DBSCAN algorithm could provide the geographical representation of the latitude and longitudes extracted and the behaviour of the user while considering a destination. It can also signify the popularity of a particular destination amongst a group of users. Table 1 compares the various works of literature carried out.

Article	**UGC**	**BP**	**TS**	**IA**	**GT**	**MC**
Shiranthika et al. (2018)	•	•		•	•	
Memon et al. (2015)	•	•			•	
Cai et al. (2018)	•	•		•	•	
Miah et al. (2017)	•	•	•	•	•	
Hasnat and Hasan (2018)	•	•	•		•	
Sun and Lee (2017)	•	•	•	•	•	
GERU et al. (2018)	•	•		•		•
Proposed Research	•	•	•	•	•	•

Table 1: Table of Comparison

UGC = User Generated Content, BP = Behaviour Patterns, TS = Time Series, IA = Image Analysis, GT = Geo Tagging, MC = Marketing Campaigns

3 Methodology

Whenever a client approaches a travel agent, it is necessary for them to have an overview of the customer preferences. The vast information over the world wide web can often confuse the customer regarding the choices. However, customer bookings and packages could be improved by having a personalised recommendation system.

3.1 Solution

In this project, the travel agent can have an overview about the customer preferences owing the social media images posted by the user. To achieve this modified CRISP-DM(Cross-industry Standard Process for Data Mining) methodology is adopted. Figure 1 elaborates the modified CRISP-DM methodology. The architecture of the system is provided in Figure 2. It consists of six modules:

1. Data Extraction
2. Computer Vision Tools
3. Categorisation
4. Geo-Location Analysis
5. Machine Learning
6. Time-series Analysis

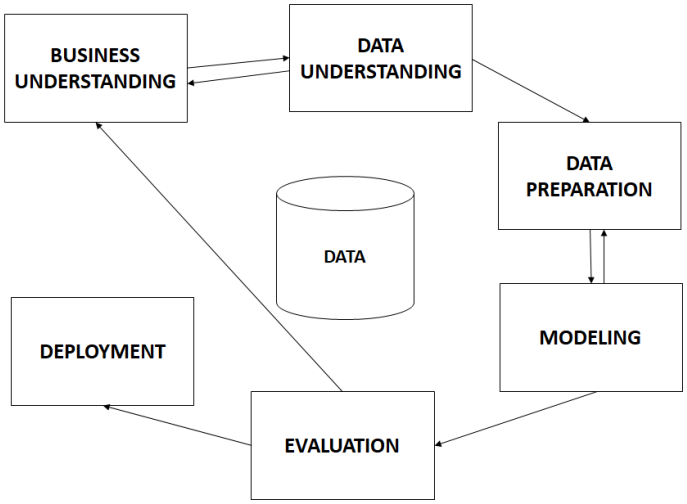


Figure 1: CRISP-DM

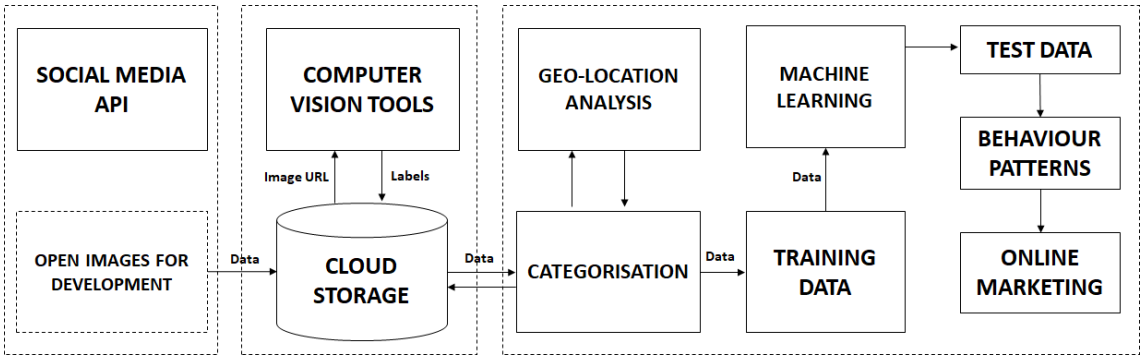


Figure 2: System Architecture

3.1.1 Data Extraction

This module is responsible for fetching data from various sources. In order to understand the preferences of the potential customer and personalising the packages, it is essential for the system to gain knowledge about that person. For facilitating this, several data sources have to be utilized. **Social media platforms require a *Production URL* of the system and hence for the purpose of demonstration of the project, an open-source image dataset from Google Open Images is being considered (please refer section 3.2)**. Social media APIs usually provide details like image URL, caption, comments, geo-location, user interactions like and comments, time, etc [6]. User travel history over a long duration of time could provide more insights about their preferences. For the geotagged images, the preferred travel location could be analysed. The data once received, the images and its meta descriptions are stored over a cloud hosting for further analysis.

3.1.2 Computer Vision Tools

Previously, extracting the features and generating labels from images was a hectic task, but over time, computer vision tools have evolved. These tools have facilitated with REST APIs. On sharing images via the APIs, information from the images can be extracted.

[6]https://developers.facebook.com/docs/instagram-api/reference/media

Amazon Rekognition is a renowned computer vision tool which is considered for the project. The image URLs from the database are shared over to Amazon Rekognition. The tool analyses the images and passes over the respective labels as the response. These labels are stored in the database for further analysis. It is also able to identify and classify objects, faces, gender, extract words and emotional states Bosch et al. (2018) present in the images. The data is stored in the database in correspondence with the analysed Image ID.

3.1.3 Categorisation

The categorisation is the process of assigning the labels to a category. In this step, low levels labels are assigned to a category which could be used for analysis. A label dictionary is being created which is linked to a particular category. A complex label dataset is reduced to such levels for a better understanding of the data. Label encoding of the categories is to be performed and respective results could be used for analysis. Every image is allocated with an independent category and henceforth the latest 10 images are utilised for denoted the user category. This information would be used for machine learning classification and pattern identification. Furthermore, this data is used for training the model.

3.1.4 Geo-Location Analysis

Platforms permit users to share their geolocation with the content they upload. This data can be extracted from the REST API can be used to understand the places travelled by a user in specific and also to understand its demand. DBSCAN is a clustering algorithm which provides a solution in this scenario Ester et al. (1996). The results consist of a cluster of points which are closely associated with each other whereas the low-density region is related to the outliers.

3.1.5 Machine Learning

For the purpose of analysing the behavioural patterns of a person, the system takes into consideration Content-Based Filtering and Collaborative Filtering techniques. In Content-Based Filtering, recommendations are made based upon the content which is uploaded by the user whereas in Collaborative Filtering recommendations are on the virtue of the individuals with similar thinking. In the training of the model, the data received from the categorisation module is being taken into consideration. This data is preprocessed as per the need of the input sequence of the model, following which the model is trained upon the following algorithms

- **KNN**: As stated by Altman (1992), the KNN algorithm has various applications in the field of classification and regression. This is a lazy learning algorithm. The algorithm is capable to distinguish the data to formulate various classes which are known as data points. New data points are formed based upon the previous outputs. As to identify the personality, KNN would suit the requirements Shiranthika et al. (2018). This algorithm cannot bear noise and hence results could be affected in its presence. Therefore, precleaning should be done with proper care.

- **K-means Clustering**: The algorithm aims to partition the data into k different clusters where every data point is belonging to a particular cluster Celebi et al.

(2013). The nearest mean or the euclidean distance is taken into consideration while allocating an observation to a cluster. Hence, the algorithm is suitable in this scenario for the classification of data.

These techniques would be essential in the classification of the potential customers' interest into Adventurous, Architecture Wanderer, Beach Person, Fashion Freak, Food Lover, Nature Explorer.

3.1.6 Time-series Analysis

Time-series analysis for the identification of behaviour patterns using social media images is a field where not much research has been conducted. **Recurrent Neural Network - Long Short Term Memory** (RNN-LSTM), can assist in predicting the next possible travel preference. this data could be obtained by analysing the change in the behavioural pattern over a period of time. The internal memory state is used to execute the flow of inputs. The impact of changes in time, weather and seasons could be studied more efficient in this module to improve the accuracy.

3.2 Datasets

For the purpose of this project, an open-source images dataset would be utilized which is available for download [7]. This dataset consists of 100,000 images in JPG format. These images are sent across the computer vision tools to extract 10 labels above 70% confidence level. A cost factor is associated with analysing these images and hence a total of 6,282 images are being examined which have produced 53,618 labels. Table 2 represents the format in which the data is stored.

	Column Name	**Description**	**Data type**
1	ID	Image ID	numerical
2	Filename	Filename of the image	text
3	UserID	ID of user associated	numerical
4	Label	Label from computer vision tools	text
5	Confidence	Confidence level	float

Table 2: Images Data Description

A corresponding Table 3 is being generated from Mockaroo [8] which represents the geo-location details. This is a piece of random information since the project is for the purpose of demonstration. The respective data could be extracted from social media APIs. Every image is then linked with the geo-location information to form a dataset which could be used for analysis.

[7] https://storage.googleapis.com/openimages/web/download.html
[8] https://www.mockaroo.com/

	Column Name	Description	Data type
1	City	City where the image was clicked	text
2	Country	Country where the image was clicked	text
3	Latitude	Location Latitude	float
4	Longitude	Location Longitude	float

Table 3: Geolocation Data Description

A random usernames dataset is being generated Mockaroo [8]. A total of 290 user-profile are generated. Table 4 elaborates the data set columns and its datatype.

	Column Name	Description	Data type
1	First Name	First Name of the user	text
2	Last Name	Last Name of the user	text
3	Email	Email address of the user	email
4	Gender	Gender of the user	male or female
5	City	City where the user stays	text
6	Country	Country where the user stays	text
7	Age	Age of the user	numerical
8	Profession	Profession of the user	text

Table 4: User Profile Data Description

Furthermore, every image is associated with a user and geolocation. The cities and geo-locations are specific to Ireland and the United Kingdom.

3.2.1 Pre-processing

Information from user profile like First Name, Last Name, Email ID are can be neglected and the model could be developed over the remaining content. As discussed in Section 3.1.3, the labels received from the computer vision tools are being assigned to a particular category; the distribution of data is listed in Table 5 and Figure 3. From this information, it is seen that the images are more related to Architecture, Adventure and Nature categories. The graphical representation highlights the percentage distribution.

	Class(Category)	Count(n)	Percentage(%)
1	Adventure	59	20.3
2	Architecture	111	38.3
3	Beach	18	6.2
4	Clothing	26	8.9
5	Food	21	7.2
6	Nature	56	19.3

Table 5: Category Distribution

Please note, the information used in the development of the model is random and open-source which is publicly available to download. The data is not fetched from any social media platform since they need a production url. Hence no copyright or ethics are exploited (section 8).

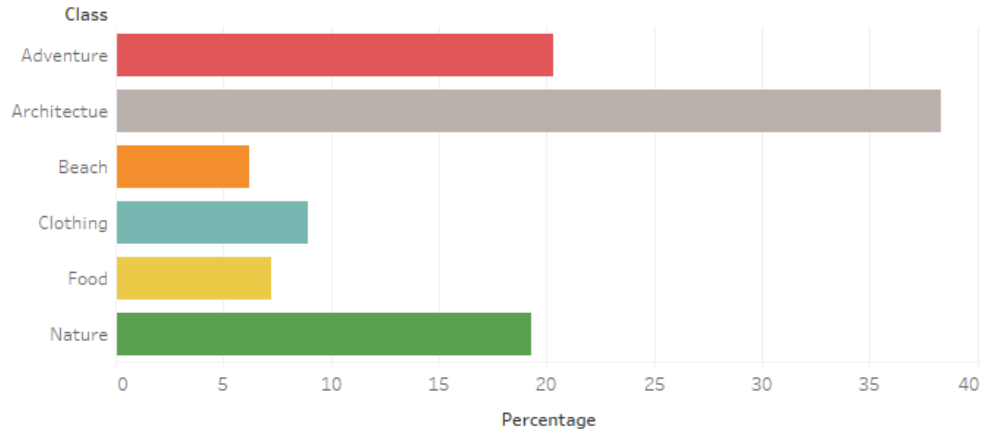


Figure 3: Category Distribution

4 Experimentation

The implementation of the system is dependent upon the following modules.

1. **Linux Server**: A Linux server over the cloud hosting is needed to process the requirement of API calls sent to the computer vision tools Yetis et al. (2016). Amazon EC2 Ubuntu 18.04 LTS SSD instance with 30GB storage. After fetching the necessary data, it is to be stored in a database for further analysis. The server is configured with Apache2, PHP 7.1.3, MySQL server.

2. **REST API**: REST web service facilitates the inter-server communication between cloud hosting and computer vision tools over the internet. The connection between social media channels could be made using this service. The information may differ from channel to channel; however, the data is stored in the database as and what needed. This web service is known for its faster processing, reliability and its availability.

3. **MySQL Database**: A MySQL database can store structured data efficiently. The data received from computer vision tools are in a structured format and MySQL database can store the data effectively in this scenario.

4. **Analytical Softwares**: Python and R are renowned analytical softwares which are used for the development of the machine learning model. These tools have enormous libraries and online documentation present. Tensorflow is used as an backend for LSTM experimentation. The results received from these model would be visualised using Tableau.

5 Results

On performing the various machine learning models, the system is able to predict the behavioural patterns of the users. The experiment was executed on various number of group of users as shown in the below tables.

5.1 KNN Algorithm

As per the work of literature, the KNN is slow learning algorithm. The Precision, Recall and F1-score improve with more amount of data. The algorithm aims to predict the users' behavioural category from the previous image information available. A group of 50, 100, 150, 200, 250 and 290 users are considered for this analysis. The results received are listed below.

	50	**100**	**150**	**200**	**250**	**290**
Precision (micro)	0.73	0.76	0.87	0.83	0.86	0.8
Precision (macro)	0.42	0.69	0.95	0.91	0.89	0.83
Precision (weighted)	0.55	0.72	0.91	0.88	0.87	0.81
Recall (micro)	0.73	0.76	0.87	0.83	0.86	0.8
Recall (macro)	0.5	0.74	0.9	0.74	0.9	0.79
Recall (weighted)	0.73	0.76	0.87	0.83	0.86	0.8
F1-score (micro)	0.73	0.76	0.87	0.83	0.86	0.8
F1-score (macro)	0.45	0.7	0.91	0.77	0.89	0.8
F1-score (weighted)	0.62	0.72	0.87	0.82	0.86	0.8

Table 6: Results of K-Nearest Neighbor Algorithm

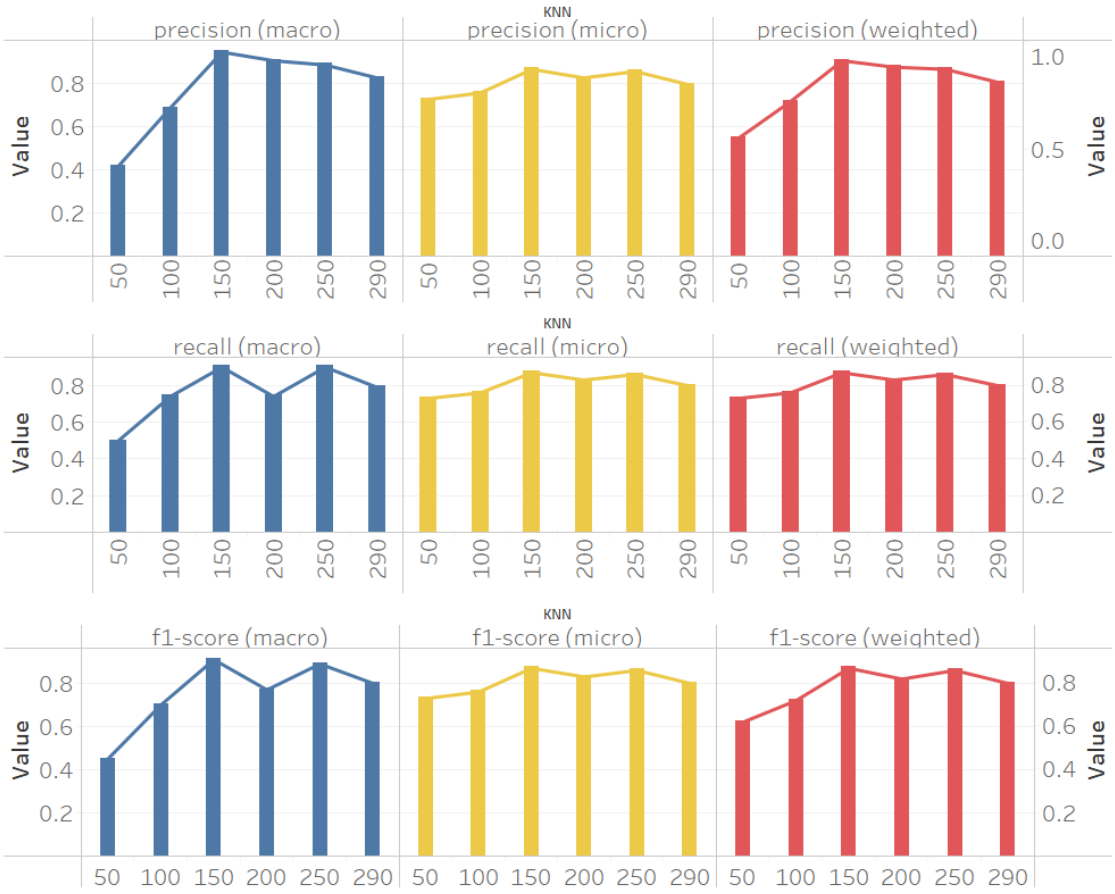


Figure 4: KNN Result Comparison

From the results represented in Table 6 and Figure 4, it can be seen that KNN is a

slow learner. The Precision, Recall and F1-Score increases with the number of users. The model is able to predict the users' behavioural patterns after 150 users which thereafter is seen constant with a percentage of around 80%. It produces an average 0.793 precision, 0.792 recall and 0.781 F1-score values.

5.2 K-Means Algorithm

K-Means clustering method forms clusters and partitions the n observations and k clusters. The experiment was performed over a dataset of 100 users aimed in forming 4 clusters. The experiment provided an accuracy of 77.33% when used the default settings and 82.66% when tuned. However, with an increasing number of observations and clusters, the model displayed a fall in the accuracy values. From the experiment, Table 7 and figure 5 provides insights of the results. This experiment was carried over 50, 100, 150, 200, 250, 290 users and the accuracy was recorded. From the results, it can be seen that the accuracy values ranging between 55% to 65%; the tuned accuracy is comparatively more on several instances.

Users	**Accuracy (%)**	**Accuracy - Tuned (%)**
50	62.5	62.5
100	62.5	64.5
150	62.5	66.66
200	57.5	56.25
250	58	58.5
290	62.06	59.91

Table 7: Accuracy of K-Means Algorithm

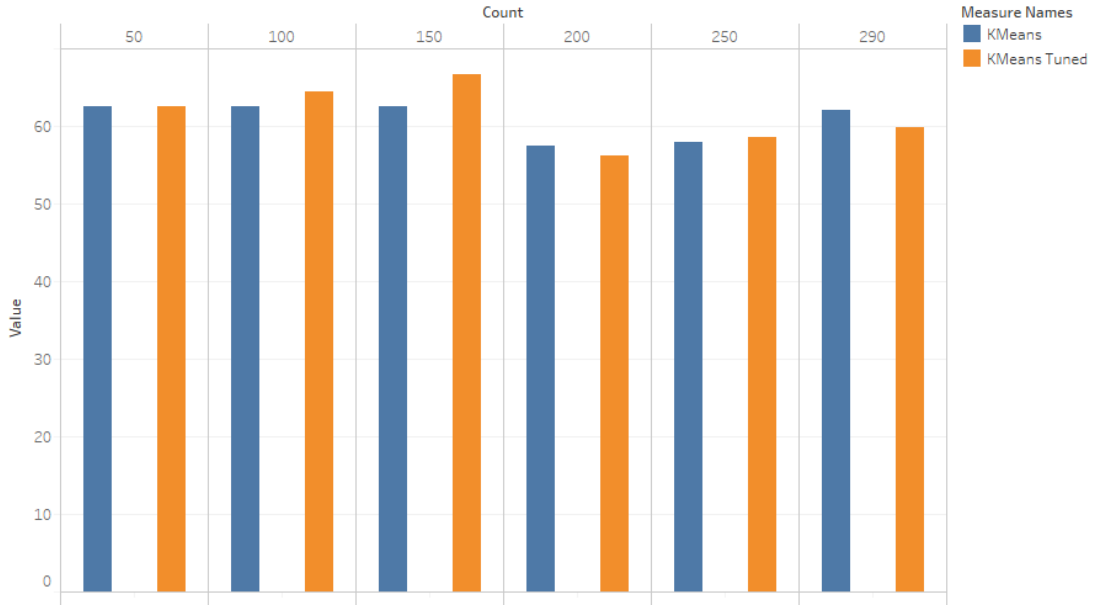


Figure 5: K-Means Accuracy Comparison

The model displayed an average accuracy of 60.843% and 61.386% with default settings and tuned-up settings. The accuracy is less as compared to the KNN algorithm

and the accuracy achieved in the previous works of literature because of the variance the observations and can improve over actual social media image data.

5.3 RNN-LSTM Algorithm

The RNN-LSTM algorithm facilitates with time-series forecasting and tries to predicts the next possible destination the user would prefer to travel. From the provided data, the model achieves a Root Mean Squared Error (RMSE) value 0.1592 and a Mean Squared Error (MSE) value of 0.02536. Figure 6 displays the predictions made by the model(*red*) against the actual values (*green*).

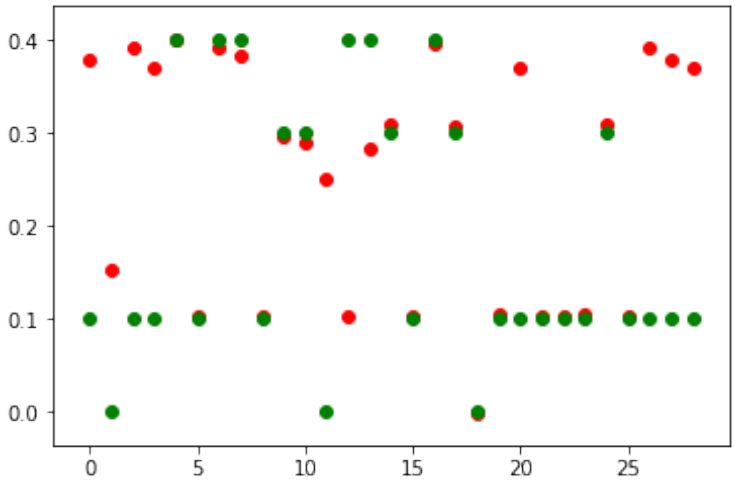


Figure 6: LSTM Predictions

6 Evaluation

The proposed system is intended to identify the behavioural pattern concerning travelling. Machine learning algorithms like KNN, K-means, LSTM and DBSCAN are being implemented owing to which the accuracy could be evaluated. The accuracy received could justify the ability of the model to predict the values in various circumstances. The following techniques could help in evaluating the strategies.

6.1 Extrapolation Testing

In this method, the users data are divided into the ratio of 75:25 as shown in Figure 7 and the machine learning make their respective predictions, thereafter the accuracy is evaluated. The accuracy is calculated for various number of users and then the overall accuracy would be extrapolated using a linear regression technique. The extrapolated accuracy would be an equitable estimate which can vary for higher number of users.

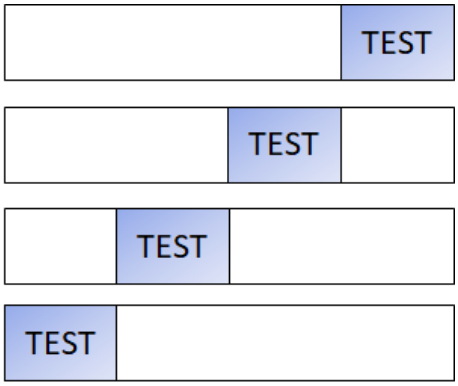


Figure 7: Data Spliting

Let A be the accuracy for n users and OAA be the Overall Average Accuracy. Then the confusion matrix would be represented as:

	Keen Interest	Not Keen Interest
Predicted Recommendation	True Positive (TP)	False Positive (FP)
Not Predicted Recommendation	False Negative (FN)	True Negative (TN)

$$A_n(\%) = \frac{(\sum_{n=1}^{10} \frac{TP_n+TN_n}{TP_n+FP_n+FN_n+TN_n})}{10} 100$$

$$A_n(\%) = (\sum_{n=1}^{10} \frac{TP_n + TN_n}{TP_n + FP_n + FN_n + TN_n}) 10$$

$$OAA(\%) = \frac{\sum_{n=1}^{k} A_n}{k}$$

These points when plotted over a two-dimension graph, the accuracy of the remaining users can be extrapolated using linear regression. A straight line through the data can provide the best fit. This line is a meer estimate and the accuracy can alter on a higher user count as the percentage cannot cross 100%. The linear regression can be represented using the following formula.

$$y(x_*) = y_{n-1} + \frac{x_* - x_{k-1}}{x_k - x_{k-1}} (y_k - y_{k-1})$$

The results of linear regression for extrapolation testing are represented in Figure 8 and Figure 9 for KNN and KMeans algorithm respectively.

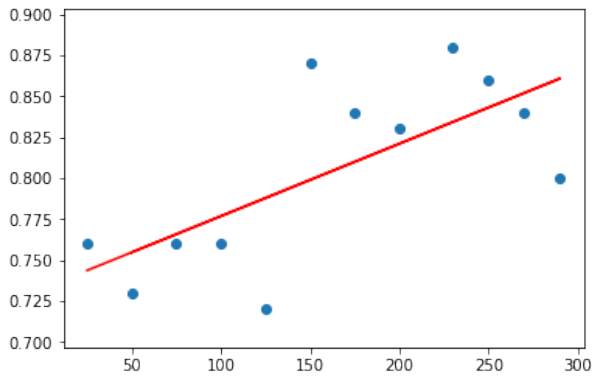


Figure 8: KNN Precision extrapolation

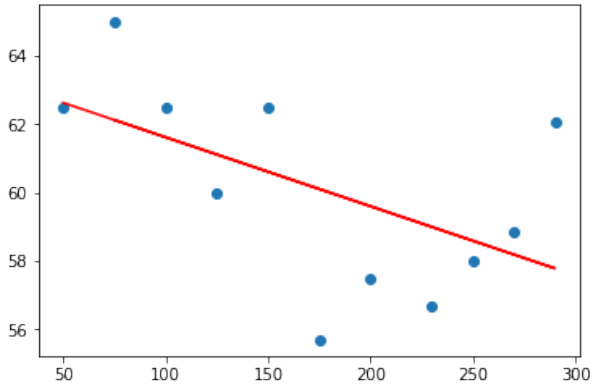


Figure 9: K-Means Accuracy extrapolation

On performing Linear Regression Extrapolation testing it is observed that the precision value of KNN (Figure 8) is showing steady growth and the accuracy of the model would improve when taken more users into consideration. The linear regression best-fit line shows an incremental trend, which can not be true if the predicted accuracy crosses 100%. However, in the case of the K-Means algorithm (Figure 9), the extrapolation testing model is suffering a continuous decline. The accuracy is observed to be variable between the range of 55 to 65%. The variance in the accuracy may be improved when analysed on actual social media user images. Owing to this evaluation, KNN would be a better algorithm when considered such a kind of analysis.

6.2 Time-Series Testing

The time series analysis of a users' profile has been carried out using RNN-LSTM and the method was able to predict the next possible travel preference as seen earlier in Figure 6. The model is user-specific and can generate personalised user predictions. It was able to produce an RMSE value of 0.1592 and MSE value of 0.02536 for a specific user. This results are user-specific and may alter for other user profiles.

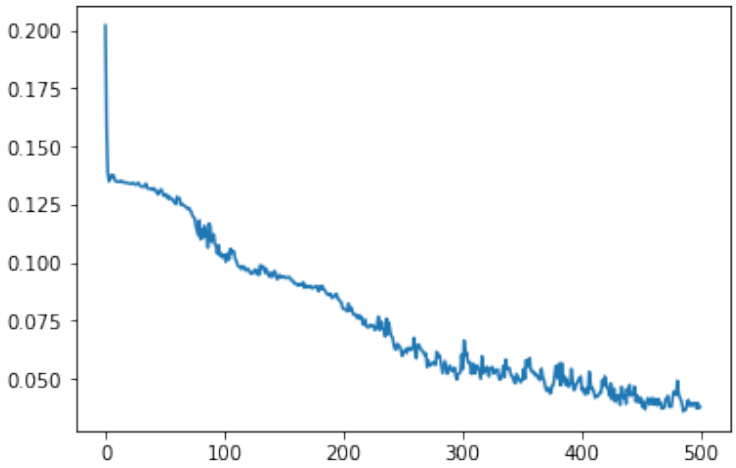


Figure 10: Error Rate with Epochs

Moreover, RNN-LSTM provides insights to the travel agent regarding the preferences of a particular client based on the time-series analysis of their social media profile. As

seen in Figure 10, the error rate has decreased drastically with every epoch alteration. The error rate is constant around 0.05 to 0.075 and seems like reducing further but those values are inferior and can be ignored. The model would have been considered pre-mature if the epochs were less than 300 as the error-rate was decreasing further. However, this value seems to be constant with higher epochs and is showing negligible improvement. This forecasting strategy could be utilised by further comparing profiles with similar thoughts and likeliness.

6.3 Geo-Location Analysis

The geo-location of the images can provide insights about the popular destination's amongst tourist; it can also produce the overview of a users' travel preferred destinations. This algorithm is useful for Knowledge Discovery in Databases(KDD). Figure 11 and Figure 12 elaborates the travelled spots by a specific user and the overall popularity of destinations respectively.

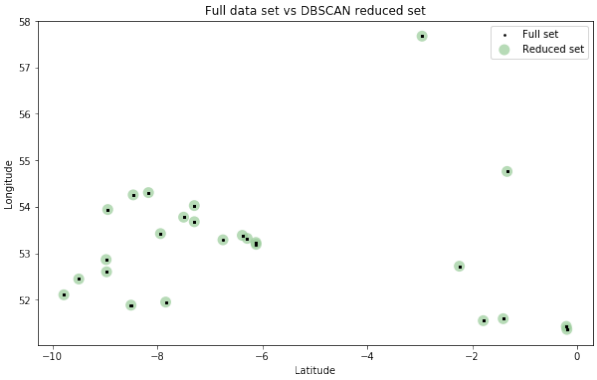


Figure 11: DBSCAN - Specific User

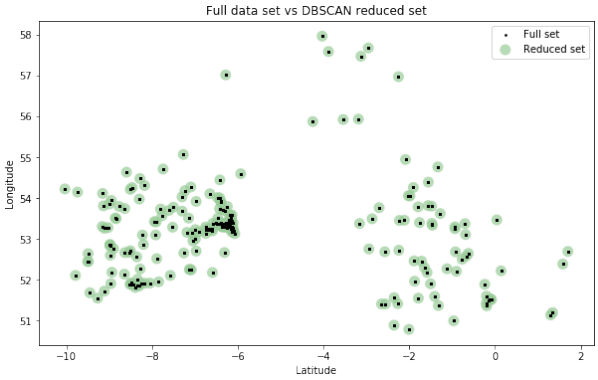


Figure 12: DBSCAN - Popular Destinations

In Figure 11, a cluster is observed and few data points which are well separated from that cluster can be seen. Similarly, in Figure 12, three distinct clusters are visible and few additional data points which are comparatively discrete. These observations can provide insights about the most user preferred destination for tourism. However, the discrete locations could be recommended people who like to travel far and explore new locations. The clusters could be further analysed for events, occasions, popularity depending upon user preferences and the package itinerary could be developed.

7 Discussion

The project work towards extracting behavioural patterns from social media images and utilising the inferences for tourism marketing. For pursuing this goal, an open images[7] dataset was used to facilitate the training of various data extraction and mining algorithms. The system discovers identical patterns present in the data and is able to classify them into categories. However, for demonstration purpose, we are focused on using open source data; unlikely more enhanced patterns could be observed on actual data. There may be differences present in the behavioural patterns based on gender and cultural backgrounds with the course of time.

The travel preferences of a potential client could be deeply studied using this system. It also symbolises, how crucial it is to identify the interests of the client and display personalised outcomes. The travel industry needs to have a wider knowledge of the users and this system ease the requirement which can lead to the improvement in conversions.

Consider a scenario in which a user hasn't tagged geo-locations over their social media content. In this case, the system could easily identify if the user is a frequent flyer depending upon the various terrains present in his social media images. Moreover, the use of Google Places API[9] could be done for listing down all the keen interest spots for a potential tourist based on travel preferences and desired location.

8 Ethics

Most social media channels need a production URL of the system for verification and thereafter they provide the application the necessary privileges. On getting the desired rights, the application can extract information from their network. For the demonstration of the system, an open dataset of images [7] was utilised. A dataset of random user information was generated using Mockaroo[8]. Therefore, **no public information was acquired or stored. Hence there is no ethical exploitation done throughout the journey of this project.**

9 Conclusion and Future Work

The research attempts in understanding behavioural patterns of the user from their social media images and utilising this information for tourism marketing. Travellers usually publish their tour images over social media channels, which are useful in providing personalised content by understanding their behavioural patterns and personality.

In this research, a total of 53,618 labels were extracted from 6,282 images. These images were assigned to 290 user profiles. Furthermore, the information was utilised for training of machine learning models like KNN, K-Means, RNN-LSTM and DBSCAN. These algorithms provided significant results and the prediction capability of the KNN model improved when tested against a larger audience. Unlikely, K-Means was not able to provide accuracy as expected; this may be a result of imbalance present in the data. RNN-LSTM algorithm strived to predict the next possible travel preference of a user. DBSCAN formed clusters and provided knowledge regarding the popular travel destination preferred. The system took into consideration collaborative filtering used in KNN and K-Means algorithms and content-based filtering used in RNN-LSTM algorithm. In future, the impact of cultural and geographical diversities over tourism can be deeply studied.

References

Ahn, M. J. and McKercher, B. (2015). The effect of cultural distance on tourism: A study of international visitors to hong kong, *Asia Pacific Journal of Tourism Research*

[9] `https://maps.googleapis.com/maps/api/place/textsearch/json?query=Letterkenny+beach&language=en&key=AIzaSyAaixYPGGAvv6tJ4hmlE772UhX4RMu8ocw`

20(1): 94–113.
URL: *https://doi.org/10.1080/10941665.2013.866586*

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician* **46**(3): 175–185.

Bosch, O. J., Revilla, M. and Paura, E. (2018). Answering mobile surveys with images: An exploration using a computer vision api, *Social Science Computer Review* p. 0894439318791515.
URL: *https://journals.sagepub.com/doi/10.1177/0894439318791515*

Cai, G., Lee, K. and Lee, I. (2018). Itinerary recommender system with semantic trajectory pattern mining from geo-tagged photos, *Expert Systems with Applications* **94**: 32 – 40.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417417307315*

Celebi, M. E., Kingravi, H. A. and Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert Systems with Applications* **40**(1): 200 – 210.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417412008767*

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise., *Kdd*, Vol. 96, pp. 226–231.

Ferwerda, B. and Tkalcic, M. (2018). Predicting users' personality from instagram pictures: Using visual and/or content features?, *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, ACM, New York, NY, USA, pp. 157–161.
URL: *http://doi.acm.org/10.1145/3209219.3209248*

Filimonau, V. and Perez, L. (2018). National culture and tourist destination choice in the uk and venezuela: an exploratory and preliminary study, *Tourism Geographies* **0**(0): 1–26.
URL: *https://doi.org/10.1080/14616688.2018.1490342*

Ge, M. and Persia, F. (2019). Factoring personalization in social media recommendations, *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pp. 344–347.
URL: *https://ieeexplore.ieee.org/document/8665624*

GERU, M., MICU, A. E., CAPATINA, A. and MICU, A. (2018). Using Artificial Intelligence on Social Media's User Generated Content for Disruptive Marketing Strategies in eCommerce, *Economics and Applied Informatics* **-**(3): 5–11.

Hasnat, M. M. and Hasan, S. (2018). Understanding tourist destination choices from geo-tagged tweets, *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3391–3396.
URL: *https://ieeexplore.ieee.org/abstract/document/8569237*

Hosseini, H., Xiao, B. and Poovendran, R. (2017). Google's cloud vision api is not robust to noise, *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 101–105.
URL: *https://ieeexplore.ieee.org/document/8260620*

Huertas, A. and Marine-Roig, E. (2016). User reactions to destination brand contents in social media, *Information Technology & Tourism* **15**(4): 291–315.
URL: *https://doi.org/10.1007/s40558-015-0045-9*

Jiang, S., Qian, X., Mei, T. and Fu, Y. (2016). Personalized travel sequence recommendation on multi-source big social media, *IEEE Transactions on Big Data* **2**(1): 43–56.

Kaosiri, Y. N., Fiol, L. J. C., Ángel Moliner Tena, M., Artola, R. M. R. and García, J. S. (2019). User-generated content sources in social media: A new approach to explore tourist satisfaction, *Journal of Travel Research* **58**(2): 253–265.
URL: *https://journals.sagepub.com/doi/10.1177/0047287517746014*

Lian, T. and Yu, C. (2019). Impacts of online images of a tourist destination on tourist travel decision, *Tourism Geographies* **0**(0): 1–30.
URL: *https://doi.org/10.1080/14616688.2019.1571094*

Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I. and Chen, G. (2015). Travel recommendation using geo-tagged photos in social media for tourist, *Wireless Personal Communications* **80**(4): 1347–1362.
URL: *https://doi.org/10.1007/s11277-014-2082-7*

Miah, S. J., Vu, H. Q., Gammack, J. and McGrath, M. (2017). A big data analytics method for tourist behaviour analysis, *Information & Management* **54**(6): 771 – 785. Smart Tourism: Traveler, Business, and Organizational Perspectives.
URL: *http://www.sciencedirect.com/science/article/pii/S0378720616303573*

Mulfari, D., Celesti, A., Fazio, M., Villari, M. and Puliafito, A. (2016). Using google cloud vision in assistive technology scenarios, *2016 IEEE Symposium on Computers and Communication (ISCC)*, pp. 214–219.

Narangajavana, Y., Fiol, L. J. C., Ángel Moliner Tena, M., Artola, R. M. R. and García, J. S. (2017). The influence of social media in creating expectations. an empirical study for a tourist destination, *Annals of Tourism Research* **65**: 60 – 70.
URL: *http://www.sciencedirect.com/science/article/pii/S0160738317300646*

Shiranthika, C., Premakumara, N., Fernando, S. and Sumathipala, S. (2018). Personalized travel spot recommendation based on unsupervised learning approach, *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pp. 229–234.
URL: *https://ieeexplore.ieee.org/document/8615533*

Sun, C.-Y. and Lee, A. J. (2017). Tour recommendations by mining photo sharing social media, *Decision Support Systems* **101**: 28 – 39.
URL: *http://www.sciencedirect.com/science/article/pii/S0167923617300982*

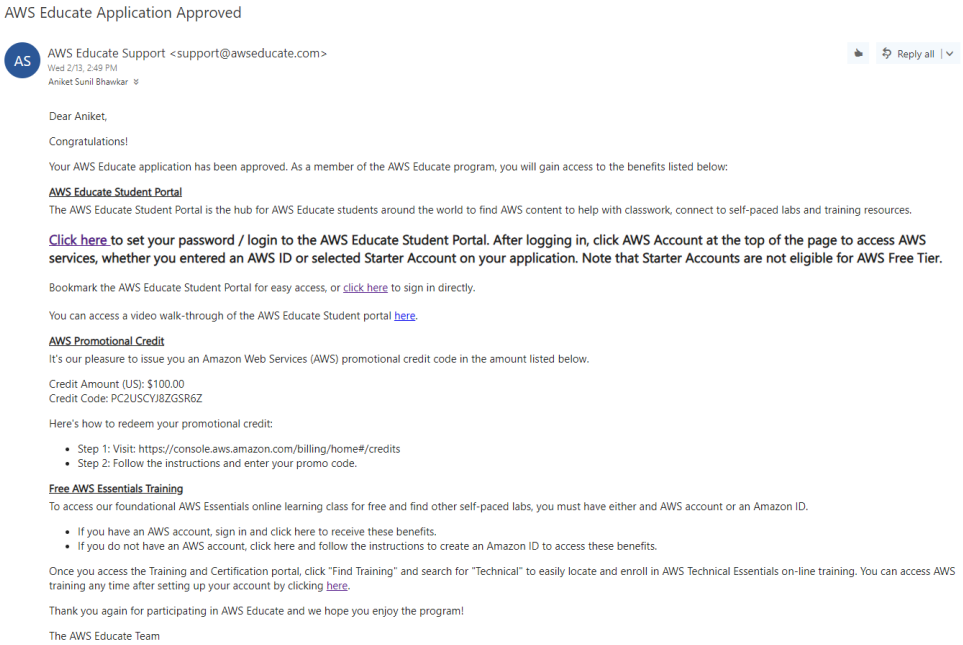
Wang, Y. and Li, B. (2015). Sentiment analysis for social media images, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1584–1591.

Yetis, Y., Sara, R. G., Erol, B. A., Kaplan, H., Akuzum, A. and Jamshidi, M. (2016). Application of big data analytics via cloud computing, *2016 World Automation Congress (WAC)*, pp. 1–5.
URL: *https://ieeexplore.ieee.org/document/7582986*

Yu, Z., Xu, H., Yang, Z. and Guo, B. (2016). Personalized travel package with multi-point-of-interest recommendation based on crowdsourced user footprints, *IEEE Transactions on Human-Machine Systems* **46**(1): 151–158.

Appendices

A Credits from AWS Educate



AWS Educate Application Approved

AS AWS Educate Support <support@awseducate.com>
Wed 2/13, 2:49 PM
Aniket Sunil Bhawkar

Reply all

Dear Aniket,

Congratulations!

Your AWS Educate application has been approved. As a member of the AWS Educate program, you will gain access to the benefits listed below:

AWS Educate Student Portal

The AWS Educate Student Portal is the hub for AWS Educate students around the world to find AWS content to help with classwork, connect to self-paced labs and training resources.

Click here to set your password / login to the AWS Educate Student Portal. After logging in, click AWS Account at the top of the page to access AWS services, whether you entered an AWS ID or selected Starter Account on your application. Note that Starter Accounts are not eligible for AWS Free Tier.

Bookmark the AWS Educate Student Portal for easy access, or click here to sign in directly.

You can access a video walk-through of the AWS Educate Student portal here.

AWS Promotional Credit

It's our pleasure to issue you an Amazon Web Services (AWS) promotional credit code in the amount listed below.

Credit Amount (US): $100.00
Credit Code: PC2USCYJ8ZGSR6Z

Here's how to redeem your promotional credit:

- Step 1: Visit: https://console.aws.amazon.com/billing/home#/credits
- Step 2: Follow the instructions and enter your promo code.

Free AWS Essentials Training

To access our foundational AWS Essentials online learning class for free and find other self-paced labs, you must have either and AWS account or an Amazon ID.

- If you have an AWS account, sign in and click here to receive these benefits.
- If you do not have an AWS account, click here and follow the instructions to create an Amazon ID to access these benefits.

Once you access the Training and Certification portal, click "Find Training" and search for "Technical" to easily locate and enroll in AWS Technical Essentials on-line training. You can access AWS training any time after setting up your account by clicking here.

Thank you again for participating in AWS Educate and we hope you enjoy the program!

The AWS Educate Team

Figure 13: Credits from AWS Educate

The research has received a scholarship of $100.00 under the AWS educate program. The amount is utilised towards AWS EC2 instance and labelling of images using Amazon Rekognition.