**Analyzing Customer Sentiment in the Tourism Industry: Insights from Reddit Data**

* Aundre Senarathna - FGS/MDA/2022/071

**Introduction**

The abundance of information on online platforms has created new opportunities for understanding and analyzing client opinions and experiences in the modern digital age. Particularly social media sites have developed into useful data sources for companies looking to understand client mood and pleasure. In order to examine customer satisfaction in the tourism business, this project will make advantage of the substantial volume of user-generated content on Reddit.[1]

By gathering and analyzing data from Reddit, this study aims to examine customer happiness in the tourism business. We can access a wide range of opinions and experiences provided by users by concentrating on Reddit, a well-known internet forum with multiple active communities. The major objectives of the research are to determine the general tone of travel-related posts, forecast customer satisfaction levels, and gather useful information to improve the travel experience.

We specified the scope of study in terms of theme focus, geographic region, and time period in order to meet our project's goal. We sought to gather concentrated and relevant data for analysis by reducing the scope. We created a crawler specifically made to collect text from posts that met our criteria in order to extract data from Reddit based on the scope we determined. To improve the analysis process, relevant metadata linked to each post was also gathered.

After gathering the data, we proceeded to pre-process it to ensure its quality and consistency. This involved removing irrelevant characters, stop words, and punctuations to focus on the essential textual content. Furthermore, we standardized the data by using a consistent format for dates and times, facilitating easier analysis. The next step was to tokenize the text data, converting it into a machine-readable format, which would allow for more efficient analysis and modeling.

Finally, we used machine learning methods to forecast consumer satisfaction levels based on the collected data, including regression, random forest, and support vector machines. These algorithms were chosen because they can handle massive amounts of textual input and produce precise predictions. We wanted to make informed judgments to improve consumer happiness in the travel sector by using machine learning techniques to find hidden patterns and relationships within the data.

**Gathering data:**

We created a crawler specifically made to extract relevant data from Reddit in order to obtain the data for our investigation. In order to verify that the data gathered was relevant and in line with our project objectives, particular criteria had to be specified during the data collection process.

There were two key components to the criteria that were utilized to collect the data. First, we restricted the data to only include articles from the previous year. This was done to make sure the information was current and to concentrate on debates and experiences that Reddit members had recently shared. We intended to capture current feelings and trends in the tourism sector by choosing a specific time frame.

During the first stage of data collection, postings that matched our criteria were gathered utilizing a list of subreddit names. Travel, SoloTravel, Backpacking, Camping and Hiking, Adventures, TravelHacks, Roadtrip, AirBnB, TravelPhotos, and TravelDeals were the subreddit names selected for data extraction. These subreddits were chosen due to their applicability to the travel sector and their potential to present a variety of user viewpoints and experiences. [2]

In addition to the initial data extraction, we improved our data collection procedure by concentrating on the word "Travel" from the subreddit. However, by focusing on particular geographic areas connected to well-known tourism destinations, we broadened our search criteria. New Zealand, Italy, Australia, Canada, Japan, Thailand, Greece, Mexico, France, Costa Rica, Maldives, Paris, Egypt, Dubai, Switzerland, Iceland, Rome, London (England), New York City, South Africa, Indonesia, Singapore, and Hong Kong were the selected target areas for data extraction. These target areas were chosen after thorough investigation and analysis, including Google searches for popular tourist spots. We want to gather a variety of travel-related discussions and experiences from Reddit users by concentrating on these regions. [3]

We took the following information from each post throughout the data collection process: title, score, author, creation date, URL and quantity of comments. The score indicates how popular the post is as measured by the number of upvotes and downvotes, while the title gives a concise description of the topic. The produced date shows when the post was published, and the author identifies the user who wrote it. The original post can be retrieved and referenced using the URL. The quantity of comments reflects how much conversation and interaction the topic has sparked.

The acquired data was recorded as CSV (Comma-Separated Values) files, a popular file format for storing structured data, to ensure simple access and analysis. This made it possible to handle data effectively and helped the analysis and modeling procedures that followed.

To ensure the relevancy, diversity, and thoroughness of the data gathered, the names of the target subreddits, localities, and data qualities were all carefully picked. We intended to acquire important insights about customer satisfaction in the travel business by embracing a wide range of viewpoints and experiences from Reddit users.

**Pre-processing:**

Following the collection of the data from Reddit, we pre-processed it to guarantee its accuracy, reliability, and suitability for study. Several actions were taken during the pre-processing stage with the goal of cleaning and standardizing the data.

First, we cleaned up the text data by removing unnecessary characters, stop words, and punctuation. This stage was required to narrow the focus on the text's core ideas and get rid of unnecessary or irrelevant data that might have influenced the analysis. By eliminating these components, we wanted to focus on the most important text and increase the precision and dependability of our analysis.[4]

We also standardized the data, by employing a common format for dates and timings. For simpler analysis and comparison across various postings and time periods, this uniformity was essential. We could reliably examine temporal trends and patterns in client satisfaction by guaranteeing a consistent format.

Tokenization was the crucial step after that in the pre-processing stage. The text data was tokenized, or divided into smaller tokens, usually words or phrases. We tokenized the text data to make it machine-readable, allowing for more effective modeling and analysis. Through this method, we were able to dissect the text in great detail and gain insightful information from each token.[5]

The cleaned text, which was produced by removing unnecessary characters, stop words, and punctuation as part of the pre-processing, was saved in a new column titled "cleaned\_text." The textual content that had been processed and used as the foundation for our study was contained in this column.

In summary, the pre-processing phase involved removing irrelevant characters, stop words, and punctuations from the text data to focus on essential content. Additionally, we standardized the data by using a consistent format for dates and times. Tokenization was applied to convert the text into a machine-readable format, allowing for efficient analysis and modeling. The resulting cleaned text was stored in a new column labeled as "cleaned\_text," and to ensure compatibility and convenience for usage in further analysis, the pre-processed data was stored as a CSV file.

**Sentiment Analysis**

A natural language processing approach called sentiment analysis is used to figure out the sentiment or opinion contained in a particular text. It entails examining the text's words and phrases to determine whether the tone is generally favorable, negative, or neutral. [6]

The SentimentIntensityAnalyzer from the nltk.sentiment.vader module was used to perform the sentiment analysis. Each text receives a sentiment score from the SentimentIntensityAnalyzer based on the presence of positive and negative terms. The four values in the sentiment score dictionary are "neg" for negative, "neu" for neutral, "pos" for positive, and "compound" for a combined score that combines the three preceding scores.[7]

The 'compound' score was used in the code to categorize a text's sentiment. The sentiment was categorized as "Positive" if the "compound" score was greater than or equal to 0.05. The sentiment was categorized as "Negative" if the "compound" score was -0.05 or less. Otherwise, the sentiment was categorized as "Neutral" if the "compound" score was between -0.05 and 0.05.

We carefully analyzed a number of characteristics that were relevant to our particular analysis in order to categorize the attitudes into categories of "Positive," "Neutral," and "Negative." We established the cutoff values of 0.05 and -0.05 through experimentation and the use of domain-specific information. These thresholds were selected to draw distinct lines between various sentiment categories. By employing these specific cutoff values, we intended to balance precision and recall, resulting in an accurate and thorough sentiment classification. The cutoffs that were chosen, carefully crafted to meet the needs of the analysis and to accurately capture the subtleties of sentiment portrayed in the text. It should be noted, however, that the choice of these cutoff values is arbitrary and can change based on the particular context and objectives of the sentiment analysis.

Below are the sentiment analysis results for the subreddit names:

|  |  |  |  |
| --- | --- | --- | --- |
| **Subreddit name** | **Negative** | **Neutral** | **Positive** |
| adventures | 11% | 48% | 42% |
| traveldeals | 13% | 47% | 41% |
| travelhacks | 12% | 55% | 33% |
| campingandhiking | 8% | 60% | 33% |
| roadtrip | 10% | 58% | 31% |
| travel | 6% | 65% | 30% |
| backpacking | 7% | 64% | 29% |
| solotravel | 24% | 48% | 27% |
| airbnb | 38% | 42% | 20% |
| travelphotos | 5% | 83% | 13% |

Travel agencies should concentrate on the top 5 subreddits with positive sentiment, according to the findings of the sentiment analysis of subreddit names, in order to increase customer satisfaction. The following subreddits have the highest percentages of positive sentiment:

* Adventures (42% positive sentiment): To appeal to the interests of members on this subreddit, travel agencies should concentrate on marketing adventure-based experiences and activities. Agencies can draw in adventure-seeking customers and increase their happiness by providing adventurous and exhilarating holiday packages.
* Travel deals (41% positive sentiment): By publicizing special offers and discounts, travel companies can take advantage of the pleasant vibe in this Reddit community. Agencies can attract people looking for affordable travel options and boost customer satisfaction by offering enticing deals and packages.
* Travel hacks (33% positive sentiment): The interest in travel tips and tricks may be shown on this subreddit. Travel companies can provide material or give services that offer helpful insider tips, travel hacks, and methods to improve the overall travel experience. Agencies can position themselves as helpful and reliable resources by offering people beneficial guidance.
* Camping and hiking (33% positive sentiment): Travel companies can appeal to outdoor adventurers who enjoy camping and hiking by providing tailored packages and itineraries. Customers will be more satisfied if you promote camping supplies, professional advice, and distinctive camping experiences to the users of this subreddit.
* Road trip (31% positive sentiment): Travel firms can design specialized road trip packages that emphasize picturesque routes, undiscovered treasures, and must-see locations. Road trip lovers' preferences might help businesses improve customer engagement and happiness.

Below are the sentiment analysis results for the top locations used in the analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| **Location** | **Negative** | **Neutral** | **Positive** |
| Switzerland | 0% | 55% | 45% |
| Costa Rica | 0% | 57% | 43% |
| Greece | 0% | 60% | 40% |
| France | 4% | 58% | 38% |
| Paris | 8% | 56% | 36% |
| Indonesia | 17% | 50% | 33% |
| Italy | 2% | 64% | 33% |
| London | 11% | 56% | 33% |
| Egypt | 0% | 69% | 31% |
| Thailand | 19% | 50% | 31% |
| Singapore | 0% | 71% | 29% |
| Canada | 0% | 75% | 25% |
| Mexico | 8% | 67% | 25% |
| Iceland | 10% | 67% | 24% |
| Australia | 0% | 80% | 20% |
| South Africa | 0% | 80% | 20% |
| Japan | 0% | 83% | 17% |
| Rome | 9% | 82% | 9% |
| Dubai | 0% | 100% | 0% |
| England | 0% | 100% | 0% |
| Hong Kong | 0% | 100% | 0% |
| New York City | 0% | 100% | 0% |
| New Zealand | 0% | 100% | 0% |

The top 10 sites/destinations that resulted from the sentiment analysis of geographic locations were determined by the percentages of positive sentiment as follows:

1. Switzerland (45% positive sentiment)
2. Costa Rica (43% positive sentiment)
3. Greece (40% positive sentiment)
4. France (38% positive sentiment)
5. Paris (36% positive sentiment)
6. Indonesia (33% positive sentiment)
7. Italy (33% positive sentiment)
8. London (33% positive sentiment)
9. Egypt (31% positive sentiment)
10. Thailand (31% positive sentiment)

When creating and promoting travel products, it is suggested that travel agents give priority to the above areas/locations. Travel agencies may increase customer happiness and boost their ratings by concentrating on locations with a positive vibe. Increased market share, revenue, and general success are expected outcomes of this strategic approach.

For example:

* Switzerland (45% positive sentiment): The promotion of Switzerland as a vacation spot renowned for its scenic landscapes, outdoor activities, and natural beauty should be a top priority for travel firms. Users looking for satisfying vacation experiences will be drawn to sights like the Swiss Alps, spotless lakes, and lovely cities.
* Costa Rica (43% positive sentiment): The promotion of Costa Rica's biodiversity, ecotourism potential, and natural beauties can be the focus of agencies. Advertising rainforests, wildlife, adventure sports, and eco-friendly travel options can help travel companies draw in eco-aware tourists and improve customer satisfaction.
* Greece (40% positive sentiment): Users on this subreddit might be interested in posts that highlight the Mediterranean cuisine, gorgeous islands, ancient ruins, and rich history of Greece. Travel firms should showcase well-known locations that provide distinctive cultural experiences and ideal island retreats, such as Athens, Santorini, and Mykonos.
* France (38% positive sentiment): Using the Eiffel Tower, the Louvre Museum, and scenic locations like Paris and Provence as their focal points, advertising firms should highlight the beauty and appeal of France. Users looking for enjoyable travel experiences in France will be interested in information about cultural events, cuisine, and French way of life.
* Paris (36% positive sentiment): By highlighting Paris' romantic atmosphere, world-class museums, and gastronomic pleasures, travel firms can profit from the positive perception of the city. Customized packages that include excursions to well-known sites like the Louvre Museum, Notre-Dame Cathedral, and walks along the Seine River can increase client satisfaction.

In summary, travel agencies should focus on experiences, goods, and locations that have generated a positive sentiment in order to increase customer happiness and strengthen their market position. Travel agencies may produce exceptional travel experiences, raise customer happiness, and ultimately flourish in the cutthroat travel industry by matching their services with the tastes and interests of the target audience.

**Pre-pandemic and post-pandemic travel industry**

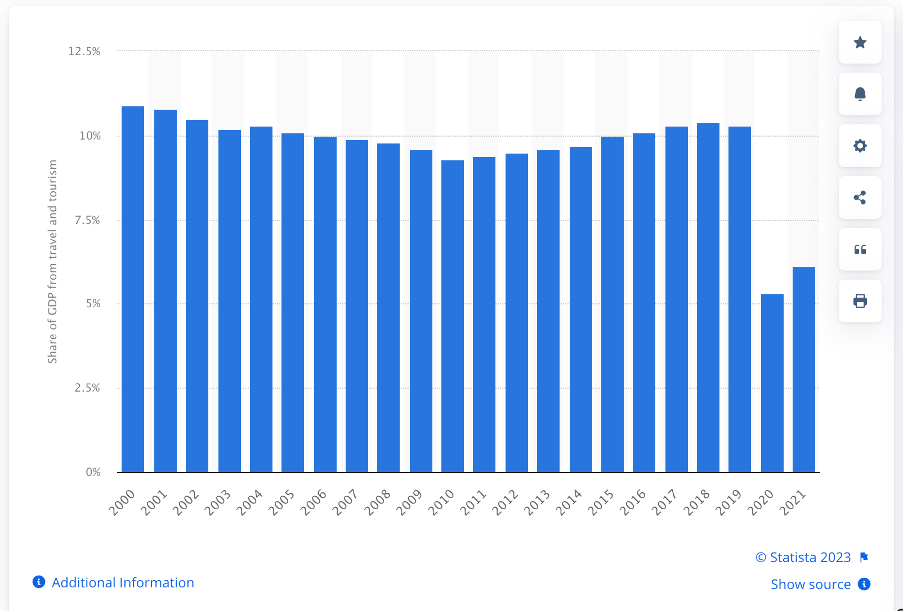
Before the COVID-19 pandemic, the travel industry was thriving and experiencing significant growth. Travel was more accessible and affordable than ever before, leading to a rise in tourism worldwide. The tourism industry was flourishing, with destinations around the world experiencing an influx of visitors. Popular tourist spots like Paris, New York City, Bangkok, and Dubai were bustling with tourists throughout the year. Many countries heavily relied on tourism for their economic growth. The sharing economy had a significant impact on the travel industry. Platforms like Airbnb and HomeAway provided alternatives to traditional hotels, allowing travelers to stay in unique accommodations and experience local culture firsthand.

The COVID-19 pandemic had a devastating impact on the travel industry, causing a significant decline in travel demand and widespread disruption across all sectors. Governments around the world implemented travel restrictions and closed their borders to control the spread of the virus. International travel came to a standstill, with countries imposing strict entry requirements, quarantine measures, and flight suspensions. Due to fear of infection and travel restrictions, travel demand plummeted. People canceled or postponed their trips, and businesses suspended non-essential travel. Hotels and accommodations faced a sharp decline in occupancy rates as travel came to a halt. Many hotels had to temporarily close or operate with limited capacity. Large gatherings, conferences, and events were canceled or postponed, affecting the business travel segment. The cancellation of major conferences and trade shows worldwide had a ripple effect on the entire travel industry, impacting airlines, hotels, and the local economy of host cities. The travel industry faced an unprecedented crisis during the pandemic, with many businesses struggling to survive. The travel industry is gradually recovering and adapting to a new normal after the COVID-19 pandemic. While the situation may vary across different regions and countries. As vaccination efforts continue and COVID-19 cases decrease in some areas, travel restrictions are being lifted, and borders are reopening. Initially, domestic and regional travel has been leading the recovery. People are more likely to opt for shorter trips closer to home, exploring their own countries or neighboring regions.

Sentiment analysis on pre-pandemic and post-pandemic

This study shows how the travel industry behaves in pre-pandemic situations and post-pandemic situations. Sentiment analysis in the travel industry involves analyzing and understanding the opinions, emotions, and attitudes expressed by individuals regarding various aspects of travel when it comes to pre-pandemic and post-pandemic situations. It helps to gauge the overall sentiment towards the industry, specific destinations, travel providers, experiences, and related topics. Sentiment analysis can be applied to customer reviews and feedback shared on websites, social media platforms, and review aggregators. Since we selected ‘Reddit’ as the source, all analytics depend on that data. According to the World Travel and Tourism Council (WTTC), the industry's direct contribution to global GDP increased steadily over the years. Here are the approximate percentages of travel and tourism's contribution to global GDP from 2015 to 2019.

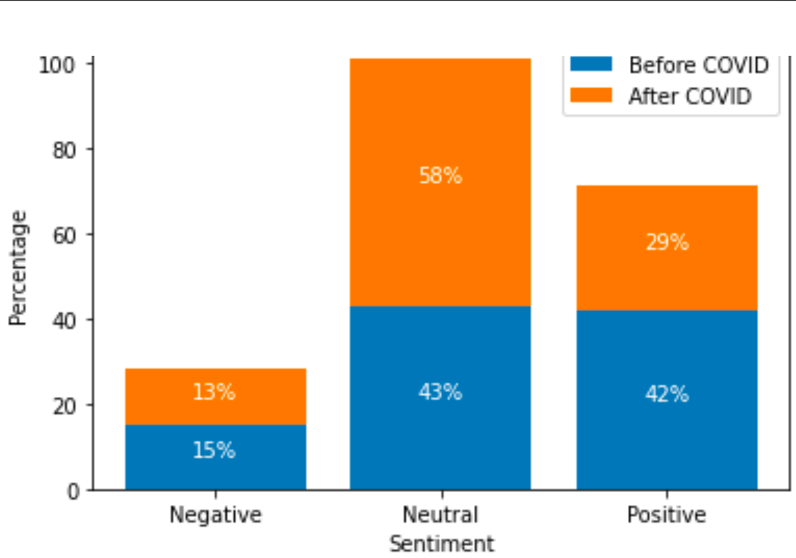
|  |  |
| --- | --- |
| Year | Contribution to Global GDP % |
| 2015 | 9.8% |
| 2016 | 10.1% |
| 2017 | 10.4% |
| 2018 | 10.6% |
| 2019 | 10.4% |



Hence, these results show that the peak of the travel industry was during the time period of 2018 to 2019 before the pandemic. Consequently, sentiment analysis was conducted during that period.

During the pandemic season, travel industry heavily went down. Hence, the result is obvious. The research focuses on the growth and peoples’ sentiment on the travel industry when it compares to the pre-pandemic situation. Selected areas in travel industry are travels, Airbnb, adventure, road trips and solo travels. These data were extracted from the globe, not a specific region or country or place. Then each object was consolidated into single sheet. All the data extracted comments that have been put in ‘Reddit’ performed sentiment analysis with nltk sentiment analysis tool. Below chart will show the percentages that have been published by the travelers under above mentioned categories

|  |  |  |  |
| --- | --- | --- | --- |
| Sentiment Analysis | | | |
|  | Negative | Neutral | Positive |
| Before covid 19 | 15% | 43% | 42% |
| After covid 19 | 13% | 58% | 29% |



Observation

It shows that there is a decrease in the negative sentiment after COVID-19 compared to the pre-pandemic period. The neutral sentiment has increased, possibly indicating a more cautious or uncertain outlook toward the travel industry. The positive sentiment has decreased, suggesting some lingering concerns or challenges faced by the industry in the post-pandemic era. Negative feedback remains the same and results in neutral and positive sentiments are significantly between pre-pandemic and post pandemic.

Recommendations

Prioritize the health and safety of travelers by implementing robust protocols and guidelines. This includes enhanced cleaning and sanitization procedures, promoting social distancing, implementing contactless solutions, and adhering to the recommendations of health authorities. Rebuild consumer confidence by transparently communicating the health and safety measures implemented by travel providers. Provide clear and up-to-date information on travel restrictions, entry requirements, and protocols. Offer flexible booking and cancellation policies to provide peace of mind to travelers.

**Prediction and Testing ML model**

Customer satisfaction prediction models are important for businesses because they can help identify areas where customer satisfaction is low and take steps to improve it. This can lead to increased customer loyalty, repeat business, and positive word-of-mouth marketing and also Testing machine learning models for accuracy is important because it ensures that the models are producing accurate results.

There are a number of different ways to create customer satisfaction prediction models. One common approach is to use machine learning algorithms to analyze historical data on customer satisfaction. This data can include things like customer surveys, feedback forms, and social media posts. Once the algorithm has been trained, it can be used to predict customer satisfaction for new customers or for customers who have not recently provided feedback.

To predict customer satisfaction, We can use machine learning algorithms such as regression or random forest or support vector machine classifier.

Sentiment scores obtained from the sentiment analysis as target variables and other features such as aspect mentions or comment length can be used as input variables.

1. Random forest -

Random forest is a machine learning algorithm that uses an ensemble of decision trees to make predictions. It is a powerful and versatile algorithm that can be used for both classification and regression tasks.

Random forest works by building multiple decision trees on bootstrapped samples of the training data. Each decision tree is trained independently, and then the predictions of the individual trees are combined to make a final prediction.

One of the advantages of random forest is that it is very good at avoiding overfitting. Overfitting occurs when a model is too closely fit to the training data, and as a result, it does not generalize well to new data. Random forest reduces the risk of overfitting by building multiple trees on different samples of the training data. [9] [10]

1. Regression analysis -

Regression analysis is a statistical technique that can be used to predict future values of a variable based on its past values. It does this by finding a linear relationship between the variable and one or more other variables. The more variables that are included in the regression model, the more accurate the predictions will be. There are many different types of regression analysis, but the most common is simple linear regression. This type of regression analysis uses a single independent variable to predict a single dependent variable. For example, you could use simple linear regression to predict the price of a house based on its square footage. Another type of regression analysis is multiple linear regression. This type of regression analysis uses multiple independent variables to predict a single dependent variable. For example, you could use multiple linear regression to predict the price of a house based on its square footage, number of bedrooms, and number of bathrooms. Regression analysis can be a powerful tool for prediction, but it is important to remember that it is not a perfect science. There will always be some error in the predictions, and the accuracy of the predictions will depend on the quality of the data that is used to train the model.[11][12]

1. Support vector machine (SVM) -

Support vector machines (SVMs) are a type of supervised learning algorithm that can be used for both classification and regression tasks. SVMs are based on the idea of finding a hyperplane that separates two classes of data points. The hyperplane is chosen such that it maximizes the margin between the two classes. This means that the distance between the hyperplane and the closest data points from each class is as large as possible.

SVMs are known for their accuracy and robustness. They can be used to solve a wide variety of problems, including image classification, text classification, and fraud detection.[13][14]

Split the data into training and testing sets -

Splitting data into training and testing sets is a crucial step in machine learning and data analysis. The purpose of this split is to evaluate the performance of a machine learning model on unseen data and to estimate its generalization ability.

Train Data set -

The training dataset is a portion of the available data that is used to train a machine learning model. It is the subset of data that the model learns from in order to make predictions or perform a specific task. The training process involves feeding the training data to the model, allowing it to analyze patterns, relationships, and dependencies within the data to optimize its internal parameters or weights [15]

Make predictions on the test set - Prediction is needed to forecast future

Calculate accuracy -

Based on the output of the above three models

**Random Forest Accuracy: : 83.8%**

The accuracy of the random forest classifier in predicting customer happiness was 83.84%. According on the specified features and sentiment scores, the model was able to classify 83.84% of the test examples correctly.

**Logistic Regression Accuracy: 81.6%**

The logistic regression classifier was 81.65% accurate in predicting customer satisfaction. This demonstrates that the model properly classified 81.65% of the test cases, which is somewhat less than the random forest model's performance.

**SVM Accuracy: 82.3%**

The support vector machine classifier had an 82.28% accuracy rate in predicting customer satisfaction. Similar to logistic regression, the SVM model fared somewhat worse than the random forest model.

According to the performance of the above three models, the Random forest model is relatively accurate for this task.

But if we consider the strength of the above three models, Random forest and SVM can effectively handle high-dimensional data and capture complex relationships between features and also Logistic regression is a simple and interpretable model that performs well when the relationship between features and the target variable is linear.

Computation through Random forest and SVM is more expensive than the Regression analysis because, They may also require more computational resources for training, especially with larger datasets.

When choosing the best model, it is important to take into account a number of criteria, including the context of predicting consumer satisfaction in the travel business. Logistic regression may be a wise option for its interpretability and simplicity of use. Random forest or SVM might be good choices if collecting complicated correlations and managing high-dimensional data are essential. The trade-off between model complexity and accuracy should be taken into account, since more complicated models may result in declining returns in accuracy improvement while raising computing costs.

**Conclusion:**

In order to analyze customer satisfaction in the tourism business, this initiative used data from Reddit. The study sought to estimate consumer satisfaction levels, identify the general tone of travel-related posts, and collect meaningful data to improve the travel experience through the collection and analysis of user-generated material.

By building a Reddit-specific crawler that concentrated on relevant subreddits and the desired regions, data was gathered. The pre-processed data was then checked for consistency and dependability. Tokenizing the text for machine reading, standardizing dates and times, and cleaning the text data were all required.

Regression, random forest, and support vector machines were used as machine learning approaches to forecast consumer satisfaction levels based on the gathered data. These algorithms were selected because they can efficiently process vast amounts of text input and produce precise predictions.

The results of the sentiment analysis showed prominent subreddits and geographic areas with favorable sentiment, recommending areas of attention for travel businesses. Travel companies may cater to Reddit users' interests and improve customer satisfaction by focusing on subreddits like Adventures, Travel Deals, and Travel Hacks. Likewise, putting Switzerland, Costa Rica, and Greece at the top of the list can assist agencies in promoting sites that inspire good feelings and draw in happy visitors.

Overall, this research emphasizes the value of utilizing user-generated material and internet platforms to learn more about customer satisfaction in the tourism sector. Travel agencies may improve the travel experience, boost client happiness, and prosper in a cutthroat industry by matching their services with customer preferences.

**References**

[1]K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, “Advances in Social Media Research: Past, Present and Future,” Information Systems Frontiers, vol. 20, no. 3, pp. 531–558, Nov. 2018, doi: <https://doi.org/10.1007/s10796-017-9810-y>.

[2]“listofsubreddits - ListOfSubreddits,” reddit. https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits/

[3]“top travel destinations - Google Search,” www.google.com. https://www.google.com/search?q=top+travel+destinations&rlz=1C1JZAP\_enLK999LK999&oq=top+travel+destinations&aqs=chrome..69i57j0i512l9.10784j0j15&sourceid=chrome&ie=UTF-8

[4]K. Ganesan, “What are Stop Words?,” Kavita Ganesan, Ph.D, Oct. 19, 2014. <https://kavita-ganesan.com/what-are-stop-words/#:~:text=Stop%20words%20are%20a%20set>

[5]A. Pai, “What is Tokenization | Tokenization In NLP,” Analytics Vidhya, May 25, 2020. <https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/>

[6]N. Barney, “What is sentiment analysis (opinion mining)? Definition from SearchBusinessAnalytics,” SearchBusinessAnalytics. <https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining#:~:text=Sentiment%20analysis%2C%20also%20referred%20to>

[7]“Sentiment Analysis,” www.nltk.org. <https://www.nltk.org/howto/sentiment.html>

[8] *Travel and tourism: share of global GDP 2021 | Statista*. (n.d.). Statista. <https://www.statista.com/statistics/1099933/travel-and-tourism-share-of-gdp/>

[9]T. Yiu, “Understanding Random Forest,” Medium, Jun. 12, 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[10]N. Donges, “A Complete Guide to the Random Forest Algorithm,” Built in, Jul. 22, 2021. [Online]. Available: https://builtin.com/data-science/random-forest-algorithm

[11]“Understanding Regression Analysis in Machine Learning,” h2o.ai. [Online]. Available: https://h2o.ai/wiki/regression/#:~:text=In%20machine%20learning%2C%20regression%20analysis.

[12]“Regression Analysis in Machine learning - Javatpoint,” www.javatpoint.com. [Online]. Available: https://www.javatpoint.com/regression-analysis-in-machine-learning

[13]R. Gandhi, “Support Vector Machine — Introduction to Machine Learning Algorithms,” Towards Data Science, Jun. 07, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[14]R. Sunil, “Understanding Support Vector Machine algorithm from examples (along with code),” Analytics Vidhya, Mar. 11, 2019. [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[15]“What Is Training Data in Machine Learning?,” MonkeyLearn Blog, Nov. 02, 2020. [Online]. Available: https://monkeylearn.com/blog/training-data/#:~:text=Training%20data%20(or%20a%20training