Sentiment Analysis on Airline Reviews and Flight Booking Time Suggestion using Six Booking Zone Concept

Karthik R[1], MuthuKrishnan R[2], Mohanakrishnan V[3]

[1]*B.E Student Dept. of CSE, Bannari Amman Institute of Technology, Erode, Tamil Nādu, India*
[2]*B.E Student Dept. of CSE, Bannari Amman Institute of Technology, Erode, Tamil Nādu, India*
[3]*B.E Student Dept. of CSE, Bannari Amman Institute of Technology, Erode, Tamil Nādu, India*

---***---

Abstract - *In this increasingly digital world, as more and more technologies emerge the connectivity between countries has increased, the demand and usage of its services have also increased. The fastest transport system to different countries is still through the air. Currently, flight ticket fares are changing programmatically like the stock market one day it's high fare and the next day it reduces drastically, even for nearby seats of the same Flight. The customers expect to travel at low fares but airlines do the exact opposite. Airlines use some strategies to charge different fares for different kinds of passengers to increase their profit. Our project aims to help the customer to book flight tickets at very low fares by use of the techniques - Optimal time to buy a ticket, six booking zones and also analyzing reviews of the airlines using natural language processing techniques to find strengths, weaknesses, and human psychology to help airlines for flight satisfaction. The task of text classification is to assign a set of predefined classes to free-text documents that can provide conceptual views for the collection of documents. The Naive Bayes classifier is a variant that is frequently used as a baseline in the process of text classification. Sentiment analysis has become an emerging trend recently due to the popularity growth of social media sites which in turn generates a humongous amount of data, due to which researchers, companies, and decision-makers has started to investigate people's thoughts and opinions in various fields. The service providers and companies are considering sentiment analysis as a valuable tool for improvement.*

Key Words: Six Booking Zone, Airline reviews, sentiment Analysis

1.INTRODUCTION

The airlines are a highly complicated industry using complex strategies to make a profit by raising ticket fares. A particular deal can change more than many times a day. These vast changes are based on what time of day or month a customer books the tickets. For customers, it's a very hard challenge to find the best possible time to book airline tickets at low fares. [1] Because they have inadequate data of the prize movements. The increasing and wide use of social media is changing how communication, information gathering, and sharing of information are being exchanged [2-4]. There is an increasing trend of companies that evaluate their performance growth by analyzing the conversations by using social media [5]. They collect the customer's opinions about their products and services. By analysing the data there has been a drastic increase in retention of costumes as the quality of services is improving. Businesses make use of the valuable data provided by Facebook, Instagram, and Twitter to track the customer's opinions of their products and their competitors.

The task of text classification is to assign a set of predefined classes to free-text documents that can provide conceptual views for the collection of documents [6]. There is a huge volume of text available in the form of digital libraries, social media, and the World Wide Web. A lot of machine learning algorithms have come into place to classify the documents automatically and therefore replace the manual classification of documents. The Naive Bayes classifier is a variant that is frequently used as a baseline in the process of text classification. It is fast and with pre-processing it gets competitive with methods like support vector machines (SVMs)[7].

Our project aims to help the customer to book flight tickets at very low fares by use of the techniques - Optimal time to buy a ticket, Six booking zones. The CheapAir.com in 2019 study recognized that there are six possible zones where customers can book tickets at low fares. First zone booking tickets 200 to 300 days second zone booking tickets 100 to 200 days in advance. In the first two zones, you have more options in seating. Third zone prime booking 20 to 100 days in advance. fourth zone 10 to 20 days in advance in the last two zones offers from airlines will be there and last two zones are lucky 0 to 10 days in this zone fares changes occur very often.

2. Literature survey:

Booking flight tickets at the best possible time was work done by (Etzioni et al)[8] The model advised the customer to Book a ticket or wait for a particular point of time. By using the price history of the airlines the model shows book tickets or wait. For analysis, the model used data mining techniques to collect over 12,000 ticket prices for 41 days of two routes. Over 61% savings was increased when compared to the normal way of Booking. This model has limitations in its analysis only for particular airlines and also only for two routes. Besides, customers can't wait for the analysis every day. Similar work to the above model was done by (William Groves, Maria Gini)[9] this model also used the same technique of book tickets based on the time" and machine learning techniques (decision tree, support vector regression) on collected ticket prices for a period of 3 months. And savings have also increased by 75%. The same authors have also done another work related to the above model [10] this time model has collected more data than previous on over 100 days ticket prices for different routes of different airlines provided customers by using REPTree classifier along with regression models. Compared to (Etzioni et al)[8] this model (William Groves, Maria Gini)[9] has handled the user request for different airlines and different routes and also increased the saving by 69%.

In 2019 a study done by cheapair.com on airline reservation service provided customers with useful information on how flight ticket fares are varying and recognized that there are six possible zones where customers can book tickets at low fares. First zone booking tickets 200 to 300 days second zone booking tickets 100 to 200 days in advance. In the first two zones, you have more options in seating. Third zone prime booking 20 to 100 days in advance. fourth zone 10 to 20 days in advance in the last two zones offers from airlines will be there and last two zones are lucky 0 to 10 days in this zone fares changes occur very often. But the drawback in all these techniques is that the customer needs to wait for the analysis signals or the customer needs to check for flights every day. That's a tedious task. In our project customers just need to provide their location details searching flights on the possible zones to get low fares and details will be provided. Customers no need to wait or search for it every day.

As the social media websites like Facebook, Twitter and Instagram have garnered a staggering number of users, these websites happen to store huge amounts of texts generated by the users of these platforms [11-13]. A large number of researchers got interested in investigating the metadata available for the purpose of the search. The researchers of [14] started to look into the gender of Twitter users. They found that many users of Twitter were using the URL section of their profile to point to their blogs which in turn provided a greater insight into their lives. A study was conducted on Sentiment Analysis of Arabic texts. The authors used machine learning techniques by using classifiers on the dataset collected. The author had investigated the effectiveness of using preprocessing methods before the text was used for sentiment classification. They concluded that the performance of the classifier had increased after the usage of preprocessing methods.

A study was conducted which looked into the ways to enhance the prediction accuracy of the stock market indicators by making use of the sentiment analysis on the Twitter data. The authors had eight emotions specific in mind out of the 755 million tweets stored and they applied a lexicon-based approach over it. The authors showed that if the straining period has increased the accuracy could be increased and they concluded that by the addition of Twitter details the accuracy did not improve drastically. On around 4,432 tweets the authors had used sentiment analysis in. Those tweets were used to collect opinions about tourism in Oman using the Concept Net. There were three lexicons namely Opinion lexicon, SentWordNet, and SentiStrength through which researchers built a sentiment lexicon. On a random basis, 80% of the data were used for training and the remaining 20% for testing. The researchers used Conceptual Semantic Sentiment Analysis and Contextual Semantic Sentiment Analysis. After applying the Naive Bayes supervised machine learning classifier they found the sentiment analysis's performance had improved by the usage of Conceptual Semantic Sentiment Analysis.

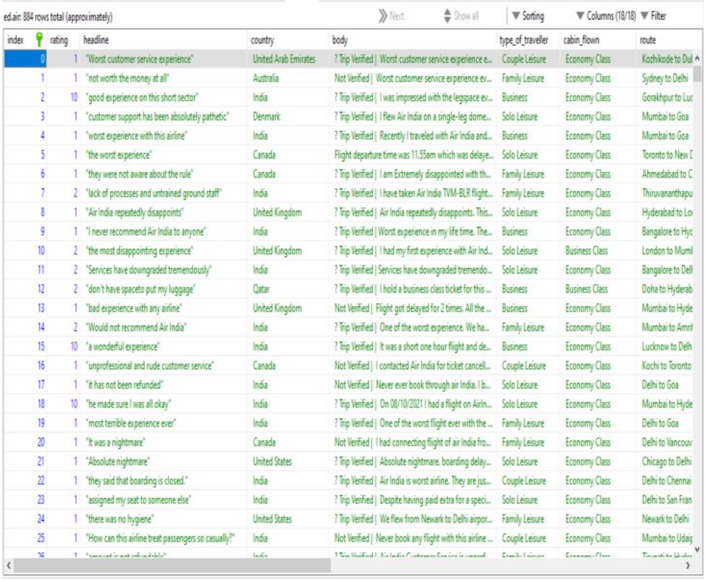
3. MAIN TEXT

3.1 Problem identification:

In previous research, the authors implemented the machine learning model to suggest to customers which time to book tickets but for particular airlines and for particular routes only which is a limitation because customers requested for different airlines and routes were not handled by that model. It also analyzed reviews only for particular airlines of historical data. In this report, we have implemented a size booking zone technique that works for different airlines and routes and which also performs sentiment analysis on live reviews.

3.2 Materials and methods:

The flight data (routes and ticket fares) are collected by using an API called Tequila which is provided by kiwi.com. For airline reviews analysis we collect the data from the airlinequality.com website by using selenium which is a web scraping module and based on our requirement it will regressively search for the reviews. In this project, we have set a review collection up to 50 pages on the site. Airline quality site which is run by Skytrax consultancy which is one of the best known for releasing airline reviews for the reliability and airline service quality evaluation. The collected data are modified into a table form for better understanding.



ed.air: 884 rows total (approximately)

index	rating	headline	country	body	type_of_traveller	cabin_flown	route
0	1	"Worst customer service experience"	United Arab Emirates	? Trip Verified | Worst customer service experience e...	Couple Leisure	Economy Class	Kozhikode to Dul
1	1	"not worth the money at all"	Australia	Not Verified | Worst customer service experience ev...	Family Leisure	Economy Class	Sydney to Delhi
2	10	"good experience on this short sector"	India	? Trip Verified | I was impressed with the legspace ev...	Business	Economy Class	Gorakhpur to Luc
3	1	"customer support has been absolutely pathetic"	Denmark	? Trip Verified | I flew Air India on a single-leg dome...	Solo Leisure	Economy Class	Mumbai to Goa
4	1	"worst experience with this airline"	India	? Trip Verified | Recently I traveled with Air India and...	Business	Economy Class	Mumbai to Goa
5	1	"the worst experience"	Canada	Flight departure time was 11.55am which was delaye...	Solo Leisure	Economy Class	Toronto to New D
6	1	"they were not aware about the rule"	Canada	? Trip Verified | I am Extremely disappointed with th...	Family Leisure	Economy Class	Ahmedabad to C
7	2	"lack of processes and untrained ground staff"	India	? Trip Verified | I have taken Air India TVM-BLR flight...	Family Leisure	Economy Class	Thiruvananthapu
8	1	"Air India repeatedly disappoints"	United Kingdom	? Trip Verified | Air India repeatedly disappoints. This...	Solo Leisure	Economy Class	Hyderabad to Lo
9	1	"I never recommend Air India to anyone"	India	? Trip Verified | Worst experience in my life time. The...	Business	Economy Class	Bangalore to Hyc
10	2	"the most disappointing experience"	United Kingdom	? Trip Verified | I had my first experience with Air Ind...	Solo Leisure	Business Class	London to Muml
11	2	"Services have downgraded tremendously"	India	? Trip Verified | Services have downgraded tremendo...	Solo Leisure	Economy Class	Bangalore to Dell
12	2	"don't have spaceto put my luggage"	Qatar	? Trip Verified | I hold a business class ticket for this ...	Business	Business Class	Doha to Hyderab
13	1	"bad experience with any airline"	United Kingdom	Not Verified | Flight got delayed for 2 times. All the ...	Business	Economy Class	Mumbai to Hyde
14	2	"Would not recommend Air India"	India	? Trip Verified | One of the worst experience. We ha...	Family Leisure	Economy Class	Mumbai to Amri
15	10	"a wonderful experience"	India	? Trip Verified | It was a short one hour flight and de...	Business	Economy Class	Lucknow to Delh
16	1	"unprofessional and rude customer service"	Canada	Not Verified | I contacted Air India for ticket cancell...	Couple Leisure	Economy Class	Kochi to Toronto
17	1	"it has not been refunded"	India	Not Verified | Never ever book through air India. I b...	Solo Leisure	Economy Class	Delhi to Goa
18	10	"he made sure I was all okay"	India	? Trip Verified | On 08/10/2021 I had a flight on Airln...	Solo Leisure	Economy Class	Mumbai to Hyde
19	1	"most terrible experience ever"	India	? Trip Verified | One of the worst flight ever with the ...	Family Leisure	Economy Class	Delhi to Goa
20	1	"It was a nightmare"	Canada	Not Verified | I had connecting flight of air India fro...	Family Leisure	Economy Class	Delhi to Vancouv
21	1	"Absolute nightmare"	United States	? Trip Verified | Absolute nightmare. boarding delay...	Solo Leisure	Economy Class	Chicago to Delhi
22	1	"they said that boarding is closed."	India	? Trip Verified | Air India is worst airline. They are jus...	Couple Leisure	Economy Class	Delhi to Chennai
23	1	"assigned my seat to someone else"	India	? Trip Verified | Despite having paid extra for a speci...	Solo Leisure	Economy Class	Delhi to San Fran
24	1	"there was no hygiene"	United States	? Trip Verified | We flew from Newark to Delhi airpor...	Family Leisure	Economy Class	Newark to Delhi
25	1	"How can this airline treat passengers so casually?"	India	Not Verified | Never book any flight with this airline ...	Couple Leisure	Economy Class	Mumbai to Udaip
26	1	"amount is not refundable"	India	? Trip Verified | Air India Customer Service is unprof...	Family Leisure	Economy Class	Tirupati to Hyder

Figure 1: Data from Airline Quality website

The modified data is processed to get meaningful data by using natural language processing. In the Data cleaning process, the raw data undergoes several stages like tokenization, stop words removal, stemming, etc.... The cleaned data will undergo classification. The classification is done by Multinomial NB which is suitable for classifying discrete features based on word counts. It requires data in integer feature counts which is done by vectorization in the data cleaning process.

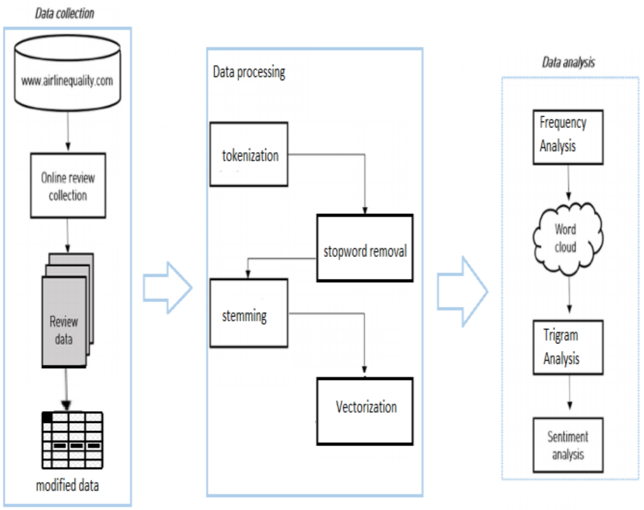


Figure 2: Data Cleaning process

3.3 Design:

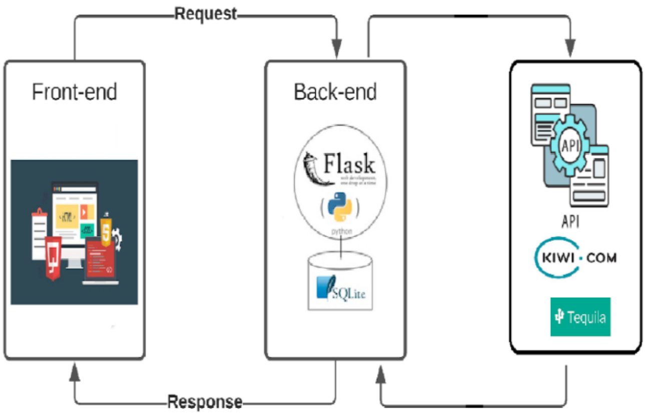


Figure 3: Overview of the Design Process

3.4 Methodology:

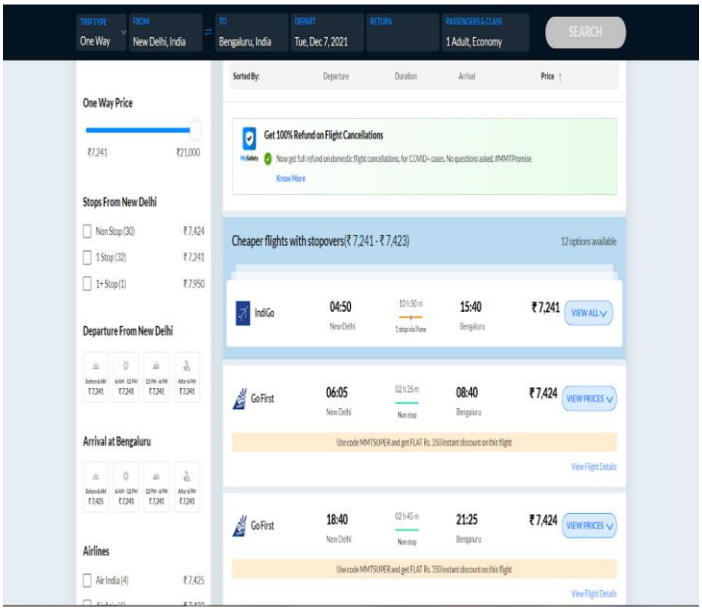
In our project, we are using the best possible timing for booking airline tickets based on the six-zone booking technique. Collecting location details (customer location, destination, number of passengers, and trip type) from customers. To get flight deals we are using a Tequila-API which is provided by Kiwi.com. In this model, the flight searching zone was set for 6 months from the time the customer request opens. This model searches for every flight for the routes requested by customers within 6 months, filter the search results which have low fares compared to normal, and show the results to the customer. The result contains on which date the ticket fares are low along with links to online travel agency sites so that customers can book tickets there itself.

Review analysis are done by using Natural Language Processing: NLP is the method of processing human language like speech and text and it is used to get meaningful data from the text data. Examples include voice assistant, sentiment analysis. Data cleaning: It is the process by which raw text is converted into clean text which involves the processes of tokenization, stemming and stop word removal. Tokenization: It involves the splitting of sentences, paragraphs or an entire text document into a number of units, words or phrases. Each small unit is called a token. Then stop word removal: Stop word removal involves the process of filtering the words whose presence in the sentence doesn't make any difference in our analysis. Then stemming involves the reduction of words into root form(removing the suffix). Then vectorization involves the process of converting text into numerical representation.

The perform classification classifies the text into different groups based on the content. Naive Bayes Algorithms involves classification where the data is grouped based on common characteristics.

4. Result and discussion:

In this work, we have created a model for the best possible timing for booking by using six possible booking zone techniques. It searches for all trips to locations requested by customers within a period of six months from the time of request filters the search results by comparing their ticket fares. And show the result along with links to book tickets.



TRIP TYPE: One Way | FROM: New Delhi, India | TO: Bengaluru, India | DEPART: Tue, Dec 7, 2021 | RETURN | PASSENGERS & CLASS: 1 Adult, Economy | SEARCH

Sorted By: Departure Duration Arrival Price ↑

One Way Price
₹7,241 ₹21,000

Get 100% Refund on Flight Cancellations
MySafety ✓ Now get full refund on domestic flight cancellations, for COVID+ cases. No questions asked. #MMTPromise
Know More

Stops From New Delhi
□ Non Stop (30) ₹7,424
□ 1 Stop (32) ₹7,241
□ 1+ Stop (1) ₹7,950

Cheaper flights with stopovers(₹ 7,241 - ₹ 7,423) 12 options available

IndiGo 04:50 New Delhi 10 h 50 m 1 stop via Pune 15:40 Bengaluru ₹ 7,241 VIEW ALL

Departure From New Delhi
Before 6 AM ₹7,241 | 6 AM - 12 PM ₹7,241 | 12 PM - 6 PM ₹7,241 | After 6 PM ₹7,241

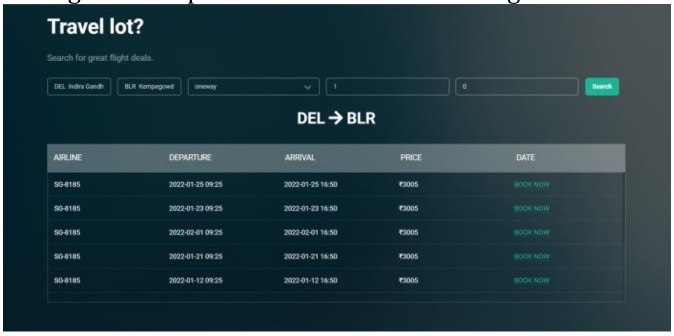
GoFirst 06:05 New Delhi 02 h 35 m Non stop 08:40 Bengaluru ₹ 7,424 VIEW PRICES
Use code MMTSUPER and get FLAT Rs. 350 instant discount on this flight
View Flight Details

Arrival at Bengaluru
Before 6 AM ₹7,425 | 6 AM - 12 PM ₹7,241 | 12 PM - 6 PM ₹7,241 | After 6 PM ₹7,241

GoFirst 18:40 New Delhi 02 h 45 m Non stop 21:25 Bengaluru ₹ 7,424 VIEW PRICES
Use code MMTSUPER and get FLAT Rs. 350 instant discount on this flight
View Flight Details

Airlines
□ Air India (4) ₹7,425

Figure 4: Trip details from online booking websites



Travel lot?

Search for great flight deals.

DEL Indira Gandh | BLR Kempegowd | oneway | 1 | 0 | Search

DEL → BLR

AIRLINE	DEPARTURE	ARRIVAL	PRICE	DATE
SG-8185	2022-01-25 09:25	2022-01-25 16:50	₹3005	BOOK NOW
SG-8185	2022-01-23 09:25	2022-01-23 16:50	₹3005	BOOK NOW
SG-8185	2022-02-01 09:25	2022-02-01 16:50	₹3005	BOOK NOW
SG-8185	2022-01-21 09:25	2022-01-21 16:50	₹3005	BOOK NOW
SG-8185	2022-01-12 09:25	2022-01-12 16:50	₹3005	BOOK NOW

Figure 5: Results from our Project

For review analysis, we collected data from airlinequality.com over 50 pages of review for a particular airline which are modified into a table format for better understanding which has around 884 rows and 18 columns

Process the collected data, remove the unnecessary words which are not useful and proceed with word cloud analysis. It is a visualization technique based on the number of occurrences of the words. Words with larger counts appear in the generated image.

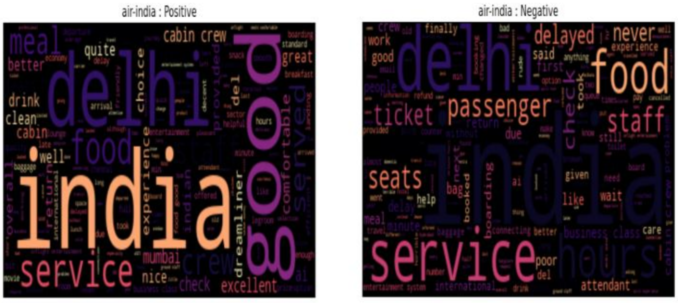


Figure 6: Word Cloud

Sentiment analysis using trigram analysis is generally used to find which word combinations are occurring often (bi-gram two words, tri-gram three words) the review data are converted into tokens (breaking the sentence into words) In this analysis the pie graph shows the most used 5 trigrams in the reviews in both positive and negative this trigram analysis are useful for airlines to improve their flight satisfaction.

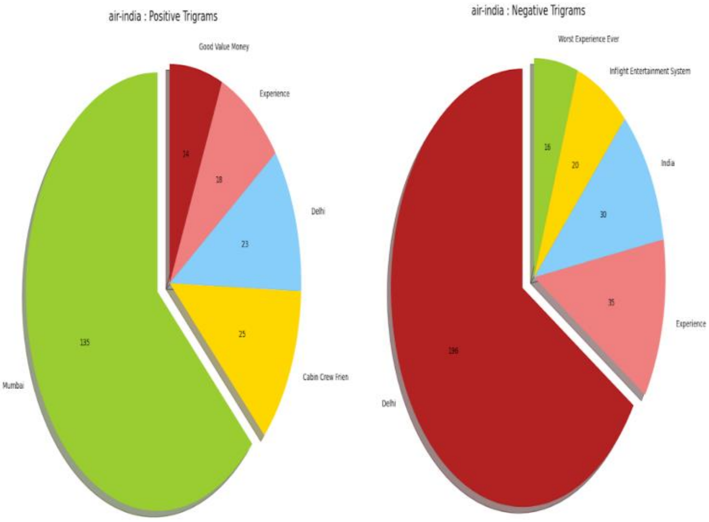


Figure 7: Trigram Analysis

Multinomial Naive Bayes is a classification algorithm used for sentiment analysis. Multinomial NB works based on the combined probability of words and classes assigned to the texts. Algorithm uses Bayes theorem to calculate the conditional probabilities.

p(A'/B) => Probability of occurrence of A when probability of B is already Occur
p(A|B) = (p(B|A) * P(A)) / P(B)

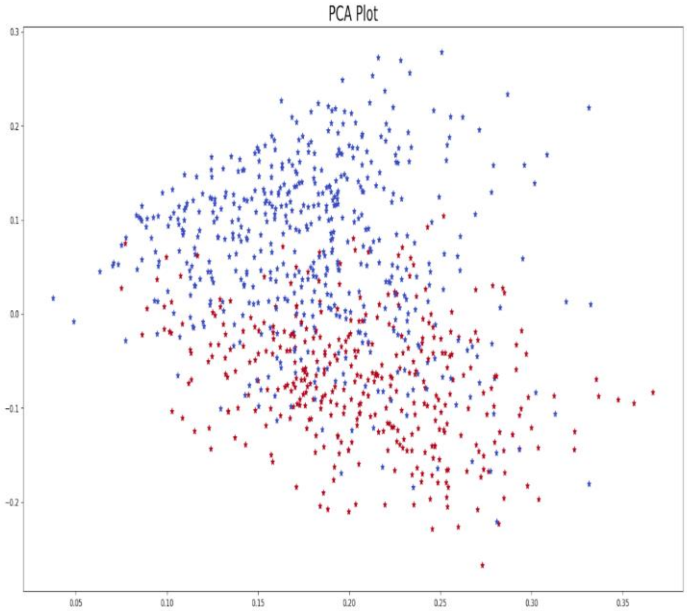


Figure 8: PCA Plot

Confusion matrix : [[139 18] [20 115]]
Accuracy score : 0.8698630136986302

5. Conclusion:

Many machine learning techniques for airline ticket fares prediction are studied. In our project, we have implemented a size booking zone technique that works for different airlines and different routes. The customers need not wait or search every day to find the lowest fare tickets. Instead, customers can just enter the location details, our model searches for the flights and gives customers the details about the flights which are available at low fares. We also perform sentiment analysis on live reviews for different airlines. Our sentiment analysis model is set for dynamic analysis where you can enter any airline. The reviews for those airlines are collected and analysis will be done. By using this model the detection of positive and negative reviews can be automated. It will be useful for airlines to improve their customer service.

REFERENCES

[1] Achyut Joshi ,Himanshu Sikaria, Tarun Devireddy (2017), "Predicting Flight Prices in India"

[2] S.A. Salloum, C. Mhamdi, B. Al Kurdi, K. Shaalan, "Factors affecting the Adoption and Meaningful Use of Social Media: A Structural Equation Modeling Approach," International Journal of Information Technology and Language Studies, 2(3), 2018.

[3] M. Alghizzawi, S.A. Salloum, M. Habes, "The role of social media in tourism marketing in Jordan," International Journal of Information Technology and Language Studies, 2(3), 2018.

[4] S.A. Salloum, W. Maqableh, C. Mhamdi, B. Al Kurdi, K. Shaalan, "Studying the Social Media Adoption by university students in the United Arab Emirates," International Journal of Information Technology and Language Studies, 2(3), 2018.

[5] F.A. Almazrouei, M. Alshurideh, B. Al Kurdi, S.A. Salloum, Social Media Impact on Business: A Systematic Review, 2021, doi:10.1007/978-3-030-58669-0_62.

[6] Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining Text Data, pp. 163-222. Springer (2012)

[7] Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R.: Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th International Conference on Machine Learning (2003)

[8] Etzioni, Oren, Rattapoom Tuchinda, Craig A. Knoblock, Alexander Yates, To buy or not to buy: mining airfare data to minimize ticket purchase price. In: 9th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, USA, August 24-27, 2003, 119-128.

[9] William Groves, Maria Gini An agent for optimizing airline ticket purchasing, in International conference on Autonomous agents and multi-agent systems International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2013

[10] William Groves, Maria Gini On optimizing airline ticket purchase timing ACM Trans. Intell. Syst. Technol. (TIST), 7 (1) (2015)

[11] C. Mhamdi, M. Al-Emran, S.A. Salloum, Text mining and analytics: A case study from news channels posts on Facebook, 2018, doi:10.1007/978-3-319-67056-0_19.

[12] A.S. Alnaser, M. Habes, M. Alghizzawi, S. Ali, "The Relation among Marketing ads, via Digital Media and mitigate (COVID-19) pandemic in Jordan The Relationship between Social Media and Academic Performance: Facebook Perspective View project Healthcare challenges during COVID-19 pandemic View project," Dspace.Urbe.University, (July), 2020.

[13] M. Alshurideh, B. Al Kurdi, S. Salloum, "Examining the Main Mobile Learning System Drivers' Effects: A Mix Empirical Examination of Both the Expectation-Confirmation Model (ECM) and the Technology Acceptance Model (TAM)," in International Conference on Advanced Intelligent Systems and Informatics, Springer: 406–417, 2019.

[14] J.D. Burger, J. Henderson, G. Kim, G. Zarrella, "Discriminating gender on Twitter," in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 1301–1309, 2011.