***Explicating Airline Passenger Reviews using Word2Vec model and Machine Learning***

V. Rathod, A. Mawandia

Amadeus Software Labs, Bangalore

***Abstract* - A large number of goods and services are being sold online. In line with the same, the practice of providing detailed online reviews as well as referring to these reviews before making purchases has gained significant traction. This work aims to unravel insights from passenger reviews of six low cost carriers from different regions of the world. The study segregates positive and negative reviews and uses Word2Vec model and Machine Learning to find natural embedding clusters. We explore the clusters of embeddings and find business insights from the same. Furthermore, we suggest an approach to classify a text as a ‘review’ or ‘not a review’ and advice use-cases in social media environment. The novelty of our work enhances the capability of an airline to identify key insights from a large bag of customer reviews.**

**Keywords—Airline Passenger Review, Word2Vec, Machine Learning, Clustering, Topic Modeling**

# Introduction

In this increasingly digitized world, a large number of goods and services are being sold online. In line with the same, the practice of providing detailed online reviews as well as referring to these reviews before making purchases has gained significant traction [1]. A large number of product reviews by verified buyers are available on ecommerce websites like Amazon and Flipkart. While a good amount of research related to machine learning techniques on product reviews by consumers on e-commerce websites is available, not much study has been done on the passenger reviews of airlines. This research attempts to unravel insights from passenger reviews of 6 key Low Cost Carriers from different regions of the world. Earlier, techniques such as topic modeling using Latent Dirchlet Allocation were used to identify topics within documents wherein each document could comprise of multiple topics in varying probabilities. This was based on probability distribution of words in these documents. This meant two documents having similar probability distribution of words would be more similar compared to those which have a non-similar probability distribution. While topic modeling techniques are good in cases where the amount of text is sufficiently large in each document and topics are clearly identifiable [2] e.g. identifying topics from news articles, it does not perform as well in situations like clustering of airline reviews as the amount of text in the reviews may be limited. Another technique which has gained immense traction in the world of automatic speech recognition, machine translation and a wide range of NLP tasks is the word embeddings or vector representations of the words in the document corpus. The word embedding computations are based on the Word2Vec model using Skip-gram model proposed by Mikolov et. al. [3]. It is an efficient method to learn high quality word embeddings from large amount of unstructured text. These word embeddings could then be used in linear combinations to arrive at meaningful linguistic patterns. For example, the result of vector calculation vec(“Spain”) – vec(“Madrid”) + vec(“France”) is closer to vec(“Paris”) than to any other word vector [4, 5 -> 8,9]. Another interesting characteristic about the Skip-gram model as highlighted by Mikolov in his research [6] is that simple vector additions could produce meaningful results. For example, vec(“Russia”) + vec(“river”) is close to vec(“Volga River”), and vec(“Germany”) + vec(“capital”) is close to vec(“Berlin”). This compositionality suggests that a non-obvious degree of language understanding can be obtained by using basic mathematical operations on the word vector representations.

We have applied the Skip-gram based Word2Vec model to generate word embeddings for the words in the airline passenger reviews. We have used these computed word vector representations in order to generate sentence vectors which would help form clusters using unsupervised machine learning techniques to derive information. The clusters so learnt would help discover insights – both positive and negative on each of the airlines. Additionally, we also propose an approach to classify a text as a “review” or “not a review” using learnt word vector representations.

The paper is organized into following sections, section 2 discusses related work on this area. Section 3 discusses the information about datasets used for research, section 4 provides the approach, section 5 presents the insights and computation results by using this model for two of the airlines, section 6 discusses the limitations of this research and provides direction for further work and section 7 concludes the paper.

# RELATED WORK

Product reviews have been of research interest for a variety of researchers including market research professionals and linguistic researchers. Some of the common tasks that have appeared in research include review classification, opinion mining, sentiment analysis, product features extraction, review summarization [4][5][6][7]. Early works from Dave et al [6] show review classification using n-gram techniques. It attempts to classify online reviews and message boards in to positive, negative and neutral classes. Similar tasks have been attempted using grouping words lexically and assigning sentiment scores [8]. Sentence bases classification approaches have used bag of sentences, naïve Bayes and SVM to achieve sentimental analysis [10]. These traditional techniques perform well in the given task of review classification / sentiment classification. Not so well in insight or summarization. Reveiws have als

The novel contribution of this work includes applying Word2Vec on online review to generate business insight.

# Information about datasets

We have collected the data from the following data sources from websites that include government, prominent websites and research reports. The dataset was available in the form of spreadsheets, tables embedded inside reports, information sections in a web page hence the challenge is to aggregate them in a consistent manner. The data mining for this research is done in a manual way.

The dataset with source details is listed in Table 1.

Table 1 . Dataset and source information

|  |  |
| --- | --- |
| **Data** | **Source Information** |
| Tier 1, 2 and 3 cities of India | Ministry of Finance, India [1] |
| Per capita Income | India Government Sources & Research Reports |
| City Population | Census India Website [12] |
| Rail Schedule | data.gov.in |
| Air Schedule | Department of Civil Aviation Website [13] |
| Train Fares | Indian Railways Website [14] |
| Air Fares | Google Flights [15] |
| Catchment area | Google Maps [16] |

# analysis approach

Rail-Fly Connectivity index (RFCI) between cities inspired by the gravity model is defined as a product of weighted values of attraction or repulsion factor and impedance factor. It is represented as:

RFCI = Weighted Attraction or Repulsion factors

**X**

(1-Weighted Impedance Factors)

Where 0 ≤ RFCI ≤1

0 ≤ Attraction or Repulsion Factors ≤1

0 ≤ Impedance Factors ≤1

The attraction factor is determined by the following factors listed in Table 3.

Table 2 . List of Attraction factors (Demographics…)

|  |  |
| --- | --- |
| **Characteristics** | **Typical Values/Measures** |
| Per Capita Income at Constant Prices | Value in Indian rupees |
| Population | Number |
| City Type | Business, Tourism or both |
| Catchment Area  (Within 200kms) | Number of Tier-2 & 3 cities within vicinity |
| Niche attributes which may cause attraction or repulsion | Include specific attributes like traveler type, purpose of travel |

The list of impedance factor is exogenous in nature like disutility aspects like distance, price, and convenience listed in Table 3. These factors may impede the Rail-Fly combination to be viable.

Table 3 . List of impedance factors (Disutility, Distance, Price…)

|  |  |
| --- | --- |
| **Characteristics** | **Details** |
| Direct Non-stop flights from origin Tier-1 city to destination Tier-2 city | Search results based on origin and destination |
| Cost differential with competition (INR) | Calculated based on differential between direct Air Fare and Air + Train Fare |
| Distance (kms) | Origin Tier-1 city to destination Tier-2 city |
| Total Travel Time including connecting time (hours) | Time taken for Fly and Rail option from Origin to Destination which includes connection time from airport to train station of connecting Tier-1 city |
| Rail-Air Schedule Alignment | Calculated value of number of trains that overlap with arrival time of flights to the connecting Tier-1 city |
| Availability of Super-Fast trains | Number of trains |

Weights calculation can be complex and need the flexibility to adjust parameters by adding and updating. The weight calculation depends on some of the following aspects below:

1. Convenience/utility factors need to be weighed based on choice models available for the passenger.
2. Passenger characteristics vary based on travel purpose.
3. Connection options from the airport to train station vary across Tier-1 cities like metro, bus, and car.
4. Existing competition within routes between Rail-Fly and direct air route connection.

To solve this problem we propose a novel approach of calculating weights using fuzzy logic. The fuzzy logic system is described in Figure 1. Fuzzy logic is generally used when the problem boundaries are unclear. In this case, the modal split is across air, rail, bus and cars between Indian cities. Fuzzy rules can aid in solving those fuzzy boundaries by defining rule base for the inference engine to calculate. The rules can be built based on the underlying characteristics using fuzzification and defuzzification process.

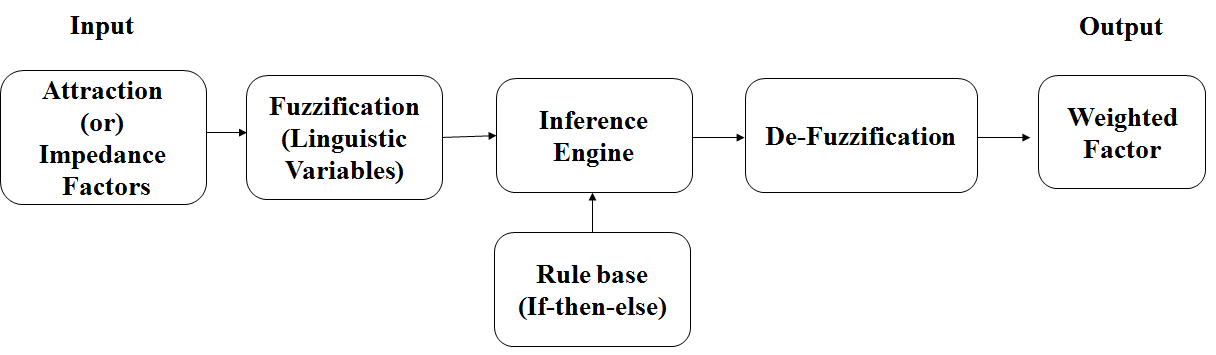


Fig.1 Fuzzy Logic Process for Weight Calculation

In the fuzzification phase, the quantitative factors are converted into linguistic variables that provide set values for fuzziness. The rules are applied to these factors by the inference engine to derive applicable linguistic value. The inference engine output is then sent to de-fuzzification for providing a crisp quantitative output weighted score for attraction and impedance factors. The value is calculated by associating a quantitative value to the linguistic value.

Table 4 provides an illustrative list of linguistic variables that can be used to convert crisp values into fuzzy values using membership functions.

Table 4 . Example of Linguistic variables

|  |  |  |
| --- | --- | --- |
| **Inputs** | **Variable** | **Fuzzy Value** |
| Population | > 3 million | High |
| Catchment Area | 3-4 | High |
| Number of Super-Fast Trains | > 20 | High |
| Number of Direct Flights | 2-3 | Low |
| Cost Difference | > 2000 | High |

Table 5 & 6 provides an illustrative list of fuzzy rules that can be applied to impedance & attraction factors respectively.

Table 5 . Example of fuzzy rule set for Impedance factors

|  |  |  |  |
| --- | --- | --- | --- |
| **Weight Rule** | **If Condition** | **If Value** | **Else Value** |
| Connection Time | High | Inconvenient | Convenient |
| Super-Fast Trains | High | Highly Convenient | Convenient |
| Cost differential | High | Good | Average |
| Direct Non-stop flights | Low | Good | Average |
| Schedule Alignment | Low | Average | Good |
| Total Travel Time | Low | Good | Average |

Table 6. Example of fuzzy rule set for attraction factors

|  |  |  |  |
| --- | --- | --- | --- |
| **Weight Rule** | **If Condition** | **If Value** | **Else Value** |
| Per capita Income | High | Favorable | Unfavorable |
| Leisure travels | High | Favorable | Unfavorable |
| Catchment Area | High | Favorable | Unfavorable |

Table 7 provides an illustrative list of defuzzification that can be applied to get the crisp value as output using centroid values.

Table 7 . Example of Defuzzification

|  |  |
| --- | --- |
| **Fuzzy Value** | **Crisp Value** |
| Favorable & Unfavorable | 0.8 & 0.1 |
| Convenient & Inconvenient | 0.7 & 0.2 |
| Good & Average | 0.8 & 0.6 |

In the next section, we will apply the above model for the southern and western cluster to derive RFCI for 2 specific routes based on service network topology [15].

# Case study of southern and western clusters

## Southern Cluster Network Analysis

Figure 2 highlights the southern network of Tier-1 and Tier-2 cities along with the distance between them. For this study, only cities within 500kms vicinity are considered. Southern cluster demonstrates the characteristics of connected hub service network topology [17] since Chennai and Bangalore are two major hub Tier-1 city within the vicinity of 300kms. The maps also list the Tier-1 (Chennai or Bangalore) city connectivity to Tier-2 cities by train.

|  |  |
| --- | --- |
| |  | | --- | |  | |

Fig.3 Southern Cluster Network

Table 8 highlights the connection possibilities from an origin Tier-1 city to the destination Tier-2 city via the Tier-1 hub city. The list is not exhaustive and there may be more combinations based on the map above.

Table 8 . Example Rail-FLY Routes

|  |  |  |
| --- | --- | --- |
| **Origin Tier-1 City** | **Connecting Tier-1 City by Air** | **Destination Tier 2 & 3 cities by Rail** |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Chennai | Coimbatore |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Chennai | Tiruchirappalli |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Chennai | Salem |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Chennai | Madurai |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Bangalore | Mangalore |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Bangalore | Mysore |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Bangalore | Coimbatore |
| Delhi/Mumbai/ Kolkata/Ahmedabad/Pune | Bangalore | Salem |

In the subsequent portions of the study, we will analyze and calculate RFCI for a specific route in the southern cluster of Mumbai-Chennai-Coimbatore. Table 9 & 10 lists the attraction and impedance factors respectively

Table 9 . Attraction factors

|  |  |
| --- | --- |
| **Characteristics** | **Value** |
| Per Capita Income | 77,975 INR [18] |
| Population | 34,58,045 |
| Type of Travel | Business, Students, Visiting Friends & Relatives |
| Catchment Area  (Within 200kms) | 4 Cities (Erode, Tirupur, Karur, and Salem) |
| Niche attributes | Garment manufacturing, machinery, pumps [18] |

Table 10 . impedance factors

|  |  |
| --- | --- |
| **Characteristics** | **Value** |
| Number of Direct Flights from Mumbai to Coimbatore | 5 |
| Cost differential with competition (Direct Air vs Air-Rail) | (537) INR |
| Distance | 1350 kms |
| Total Travel Time (Rail-Fly + Connection) | 11 hours |
| Rail-Air Schedule Alignment (Matching trains to flight arrivals) | 30+ trains |
| Availability of Super-Fast trains | 24 |

## Results and Discussion

From the fuzzy logic calculation done:

* Weighted Scored for Attraction Factors = 0.84
* Weighted Score for Impedance Factors = 0.76
* RFCI for Mumbai-Chennai-Coimbatore = 0.2

The values are derived by applying the rules in a spreadsheet and computing using the formulas available.

From the RFCI we can infer that this route has below average connectivity index considering high impedance factors related to utility and convenience. This route has a direct competition with Air and considerable improvements needed in reducing the overall travel time of train between Chennai to Coimbatore to make this a viable option. The rail route is quite popular with 24 odd express trains that connect catchment area of Salem, Erode, and Tirupur hence this route potentially can be considered for high-speed trains considering the attraction factors and in short term decreasing overall travel time might yield results.

## Western Cluster Network Analysis

Figure 3 highlights the southern network of Tier-1 and Tier-2 cities along with the distance between them. For this study, only cities within 500kms vicinity are considered. Western cluster demonstrates the characteristics of corridor hub service network topology since Pune, Mumbai, Ahmedabad, Vadodara, and Surat provides a connectivity of the corridor.

|  |
| --- |
|  |

Fig.3 Western Cluster Network (Population listed)

Table 11 highlights the connection possibilities from an origin Tier-1 city to the destination Tier-2 city via the Tier-1 hub city in the western cluster. The list is not exhaustive and there may be more combinations based on the map above.

Table 11 .Example Rail-fly routes

|  |  |  |
| --- | --- | --- |
| **Origin Tier-1 City** | **Connecting Tier-1 City by Air** | **Destination Tier 2 & 3 cities by Rail** |
| Delhi/Kolkata/Bangalore | Mumbai | Surat |
| Delhi/Kolkata/Bangalore | Mumbai | Nashik |
| Delhi/Kolkata/Bangalore | Mumbai | Aurangabad |
| Delhi/Kolkata/Bangalore | Ahmedabad | Vadodara |
| Delhi/Kolkata/Bangalore | Ahmedabad | Jamnagar |
| Delhi/Kolkata/Bangalore | Ahmedabad | Rajkot |

In the subsequent portions of the study, we will analyze and calculate RFCI for a specific route in the western cluster of Delhi-Mumbai-Surat. Table 12 & 13 lists the attraction and impedance factors respectively.

Table 12 . Attraction factors

|  |  |
| --- | --- |
| **Characteristics** | **Value** |
| Per Capita Income | 52,230 [19] |
| Population | 45,91,246 |
| Type of Travel | Business, Leisure, Visiting Friends & Relatives |
| Catchment Area  (Within 200kms) | Vadodara |
| Niche Attributes | Ornament industry like diamond merchants, machine tools, textiles, petroleum related [20] |

Table 13 . impedance factors

|  |  |
| --- | --- |
| **Characteristics** | **Value** |
| Number of Direct Flights from Delhi to Surat | 2 |
| Cost differential with competition (Direct Air vs Air-Rail) | 2551 INR |
| Distance | 1,154 kms |
| Total Travel Time (Rail-Fly + Connection) | 6 hrs |
| Rail-Air Schedule Alignment (Matching trains to flight arrivals) | 113 |
| Availability of Super-Fast trains | 51 |

## Results and Discussion

From the fuzzy logic calculation done:

* Weighted Scored for Attraction Factors = 0.76
* Weighted Score for Impedance Factors = 0.24
* RFCI for Delhi-Mumbai-Surat = 0.58

From the RFCI we can infer that this route has promising connectivity index considering low impedance factors related to utility and convenience. This route has almost no direct competition with Air and the time taken is within 6 hours. The rail route is quite popular with 51 odd express trains that connect catchment area of Vadodara and to an extent Ahmedabad hence this route might be considered for Rail-Fly option easily. Also, the corridor type network makes this route very attractive in the western cluster with all the cities having good rail and air connectivity.

Subsequently, the research can study the Northern cluster which is based on hub type with Delhi as Tier-1 city serving a lot of Tier-2 cities and the Eastern cluster which is based on Kolkata as a hub. In the next section, we will discuss the limitations and further work for this research.

# Limitations and directions for future work

This research work has the following limitations:

1. It is inspired by the gravity model with empirical values applied thru approximation techniques. It can be improved for more precision.
2. Testing the fuzzy logic rule made in real world scenario is difficult since the vastness of travel segments and clusters are concerned.
3. Bus and Car mode options are not considered for the study. India has an extensive road network across these cities and proportion of population utilize them with the modernization of highways that provides more driving speed.
4. The datasets are gathered from multiple data sources across time horizon, there may be some errors associated with the same.
5. There is a challenge to collect the data from a single source of truth hence aggregating, validating and correlating them isn’t completely possible.

The current work can be extended further to the following areas:

1. Apply the same model for Northern and Easter cluster routes.
2. Apply the model of all possible combination of routes within the same cluster and perform a sensitivity analysis of various factors.
3. In this paper, we had calculated the fuzzy logic results using spreadsheets but inference engine can be completely automated with fuzzy rule calculations which can handle all the route combinations. There are tools available like JFuzzyLogic [21] that can be used for this purpose.
4. Identify viable routes for high-speed trains and discuss with policy makers of Rail & Civil aviation for a potential partnership.
5. Understand the last mile problems like connection times and how to make passengers access transport services reach faster between airport and train stations.
6. Provide insights to further strategy development under the umbrella of smart cities program that actively promote multi-modal options
7. Research reports predict meteoric rise on affluence and the growth of various Indian cities [22] [23] which can be studied further in the context of attractions factors.
8. This research work can be augmented to study the policy and coordination needed across Government ministries and departments for identifying the gaps that need to be addressed.

# conclusion

Multi-modal travel option is a very attractive value proposition for Indian passengers and travel market in general. It solves the problem of connectivity with well-established train routes and booming air travel with low-cost carriers. Indian passenger will be the ultimate beneficiary if some of the multi-modal routes are seriously considered and actively implemented for reducing congestion across routes, more importantly allowing passenger is provided multiple choices for travel. Fuzzy logic, gravity model, and service network design are very relevant to Indian context considering the demand predictions across modes of transport, growth pattern across cities and how they are connected respectively. This work on RFCI can be easily extended further for policy planning, analysis and implementations for multi-modal transport requirements for current and future needs.

##### References

1. **Online review increase reference TBD**
2. **Online review increase reference TBD**
3. Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).
4. Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." Mining text data. Springer, Boston, MA, 2012. 415-463.
5. Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." Proceedings of the 12th international conference on World Wide Web. ACM, 2003.
6. Zhang, Dongwen, et al. "Chinese comments sentiment classification based on word2vec and SVMperf." Expert Systems with Applications 42.4 (2015): 1857-1863.
7. Mason, Rebecca, et al. "Microsummarization of Online Reviews: An Experimental Study." AAAI. 2016.
8. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." Lrec. Vol. 10. No. 2010. 2010.
9. Ohana, Bruno, and Brendan Tierney. "Sentiment classification of reviews using SentiWordNet." (2009).
10. Baharudin, Bharum. "Sentence based sentiment classification from online customer reviews." Proceedings of the 8th International Conference on Frontiers of Information Technology. ACM, 2010.
11. Ministry of Finance, Government of India, <http://ccis.nic.in/WriteReadData/CircularPortal/D2/D02ser/11016_2_2008-AIS-II.pdf>
12. BIAL Website, <http://www.bengaluruairport.com/transport/flybus.jspx>
13. India's Urban Awakening: Building Inclusive Cities, Sustaining Economic Growth, April 2010,