(7135CEM)
Modelling and Optimisation Under Uncertainty

Hotel Review Sentiment Analysis using LDA, BERT and LSA Topic Modelling Techniques.

Module Leader : Dr Alireza Daneshkhah

Student Name: Priyanka Nelli Komath
Student ID: 11979142

TASK 1:

Abstract: This paper discusses the various topic modelling for sentimental analysis of hotel review. Different techniques used here are LDA (Latent Dirichlet Allocation), LSA (Latent Semantic Analysis) and BERT. After the study the results were compared using the topic clustering, and we came to a conclusion that the best performing technique among the 3 is LDA and BERT. There are several visualisations done on this study for the easement of comparing and explaining the methods used and how it is done. This research can act as a first step for the hotel review sentiment analysis which will allow both the user and the company to act upon the result of the comparison.

Keywords : Topic Modelling, LDA, LSA, BERT, Sentiment Analysis, Hotel Review, Trip Advisor, Latent Semantic Analysis, Latent Dirichlet Allocation

Introduction

With the invention of the internet and implementation of websites has made a significant growth in online business. The number of people travelling to multiple different places has grown in this era and they depend on web applications or web sites for online hotel booking. One of the major contributors in the online hotel business is tripadvisor. With every booking the website allows the users to provide rating and the feedback of their experience. Analysing these customer experiences is essential for both consumers and merchants. Review helps the users to choose the right stay, and it helps the merchants to check the good feedback and also determine the needs of their clients and to provide them the expected service in order to keep the consumer happy.

Tripadvisor is widely used by users across the globe. As this is a popular online booking site, the number of reviews will also be huge which will make it very difficult for the users to categorise and analyse this data. The use of topic modelling techniques makes it simple to classify and organise large amounts of data, and makes it easy to categorise the reviews to good, neutral or bad. Analysis of user sentiments using topic modelling will allow the merchants to prioritise and user feedback and address them.

Topic modelling is a statistical modelling technique which can be used to identify the general "topics" that appear in a group of texts. It is an unsupervised machine learning technique and is carried out by scanning a collection of documents, identifying word and phrase patterns within them, and then automatically clustering word groups and related phrases that best describe a collection of documents. With this knowledge, we can readily determine what each group of texts is discussing. As this is an unsupervised technique it does not need training. This is the simple and quick technique to begin the data analysis process.

The goal of this project is to quantify and categorise reviews and other written content in order to help hotel providers filter out unfavourable comments and take corrective action. Additionally, it aids

consumers in choosing the top hotels based on reviews. In this research, I will be using multiple topic modelling techniques for the hotel review sentiment analysis. I am using topic modelling techniques such as LDA, LSA, and BERT to extract sentiments in the review.

One of the topic modelling techniques that analysts employ most frequently is latent semantic analysis (LSA). It is founded on the so-called distributional hypothesis, which claims that by examining the circumstances in which words are used, it is possible to understand their semantics. In other words, according to this theory, two words will have similar semantics if they frequently appear in comparable circumstances.

Transformer is an attention mechanism that learns the contextual relationships between words (or subwords) in a text and is used by BERT. Transformer's basic design consists of two independent mechanisms: an encoder that reads the text input and a decoder that generates a job prediction. Only the encoder mechanism is required because BERT's aim is to produce a language model.

The distributional hypothesis and the statistical mixing hypothesis, according to which a statistical distribution can be determined are the foundations of both Latent Dirichlet Allocation (LDA) and LSA. Each document in our corpus is mapped using LDA to a collection of topics that encompass the majority of its words.

Dataset

The dataset used for this research is taken from the online kaggle repository which was originally crawled from the trip advisor website. The data set consists of 2 columns “Review text” and “Review Rating(Star)”. The first column contains the review comment from users and the second column contains the review rating. The review rating has 5 different values from 1 to 5 where 5 being the highest and the 1 being the lowest review. The chosen dataset has 20491 rows, and it has no null or missing values.Information of the dataset with data type details are showcased in *Fig 1,* and sample data from the dataset is shown in *Fig 2.*

Link to the dataset is : https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20491 entries, 0 to 20490
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Review  20491 non-null  object
 1   Rating  20491 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 320.3+ KB
```

Fig 1: Dataset information

```
data.head()
```

	Review	Rating
0	nice hotel expensive parking got good deal sta...	4
1	ok nothing special charge diamond member hilto...	2
2	nice rooms not 4* experience hotel monaco seat...	3
3	unique, great stay, wonderful time hotel monac...	5
4	great stay great stay, went seahawk game aweso...	5

Fig 2: Sample data from dataset

Methodology

This research study discusses various modelling techniques used for analysing hotel reviews.

1. LDA (Latent Dirichlet Allocation)
2. LSA (Latent Semantic Analysis)
3. BERT

(1) LDA - Latent Dirichlet Allocation

Latent Dirichlet Allocation which is abbreviated to LDA is one of the most popular topic modelling techniques used for extracting terms from the group of written texts. Latent refers to something that is there but not fully formed. So it refers to something that is concealed or hidden. As these extracted terms are hidden, it refers to the terms 'Latent' in LDA.
LDA considers a group of topics as documents and a group of tokens as topics, it discovers hidden topics from the document. It accomplishes this by assuming potential subjects from the words used in the papers. To accomplish this, it makes use of Dirichlet distributions and a generative probabilistic model. Based on a Bayesian framework, LDA makes the inferences. As a result, the model is able to infer topics from the observable data by using conditional probabilities. In order to comprehend the observed data, a generative probabilistic model first generates data that is comparable to it. A generative model learns the key characteristics that define the data by learning how to generate observed data, which is a sophisticated method of data analysis that goes beyond simple description.
Fig 3 explains how the topic modelling works and the steps involved.

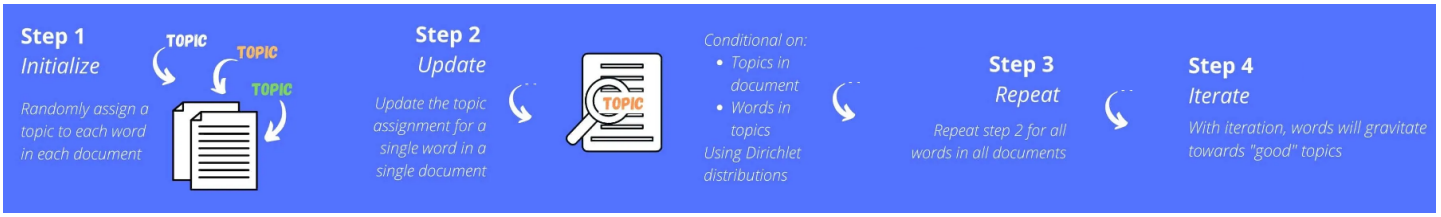


Fig 3 : Steps involved in topic modelling

(2) LDA - Latent Semantic Analysis

One of the fundamental methods used in topic modelling is LSA, or Latent Semantic Analysis. The main concept is to attempt to divide a matrix of documents and phrases into two independent matrices, those are (1) document topic matrix and (2) topic term matrix. As a result, matrix decomposition on the document-term matrix using singular value decomposition is included in the learning of LSA for latent topics. It is frequently applied as a noise- or dimension-reduction approach.

The LSA model replaces the raw counts in the document-term matrix with F-IDF scores. The idea is that words with similar importance would recur in texts that are related to each other. Utilizing shortened SVD, dimensionality reduction is accomplished (singular value decomposition). Because it depends on SVD, it takes a long time to process and is challenging to keep up with new document papers.

Fig 4 explains the workflow of Latent Semantic Analysis model

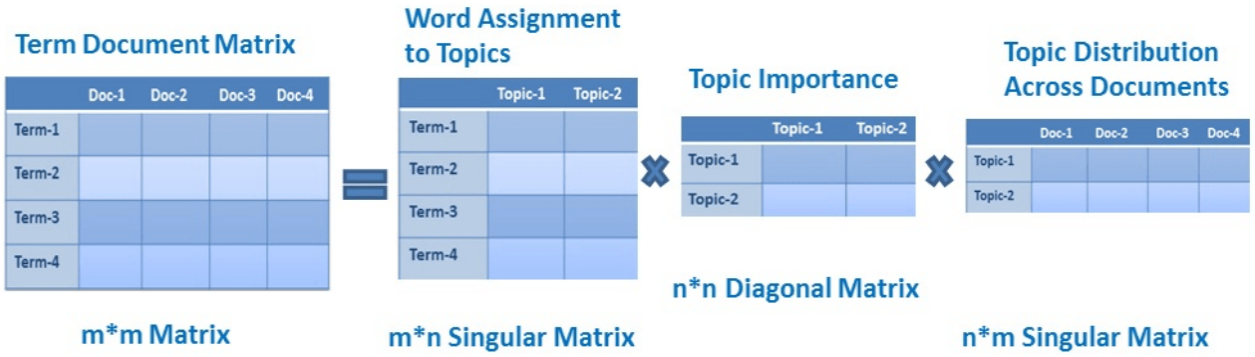


Fig 4 : LSA workflow

(3) BERT (Bidirectional Encoder Representations from Transformers)

BERT is one of the most advanced topic modelling techniques. In general, BERT is a deep neural network made up of several transformer layers. To efficiently build a language model over the corpus, the BERT model is pre-trained using a sizable corpus. With the use of transformers and the c-TF-IDF, the topic modelling technique BERT uses dense clusters to produce topics that are simple to understand while preserving key terms from the topic descriptions.

One of the most cutting-edge approaches to language topic modelling available today, BERT makes use of the improved language capabilities of these transformer models and combines them with other ML magic, such as UMAP and HDBSCAN.

Experimental Setup

Implementation of this study is done in a Windows 10 system, and I am using python programming language, and google colab IDE for implementing the written code.

As a first step, I have done the data analysis using the chosen data set. Then necessary preprocessing steps were taken to make the data ready to implement various topic modelling techniques.

(A) Data Analysis

Data analysis is carried out to discover the characteristics and instances for improved data preparation. The chosen dataset is in ".csv" format. I have uploaded the csv file into google drive and mounted the drive in the code to make the csv file accessible for implementation.
After the analysis, it is found that there are no missing or null values in the chosen dataset, hence no data adjustment or removal of duplicate values are not required.

There are two columns in the dataset "Review" and "Rating". There are 20491 reviews captured and are used for this study. "Review" is a text column and "Rating" is captured as a number value. Rating has 5 different values from 1 to 5. Figure 5 showcases the most common words in the dataset which is represented as a word cloud.

Fig 5: Word cloud representation of the dataset

From the word cloud we can see that the words “hotel”, “room”, “resort” and “time” are the dominant words in the review sentences.

After grouping the data based on the rating given by the users, the same has been plotted with the count of each rating (represented in figure 6) and their percentage (represented in figure 7).

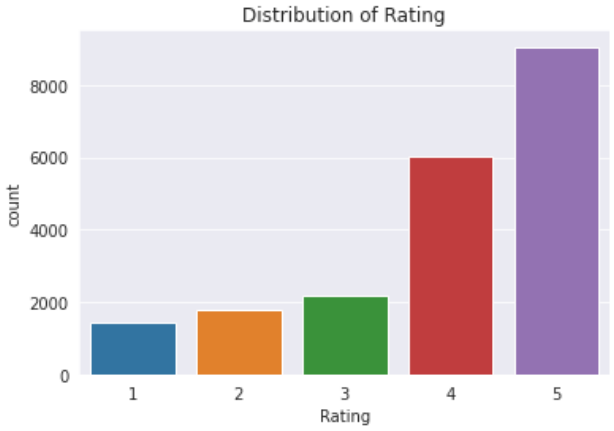


Fig 6: Distribution of reviews based on rating

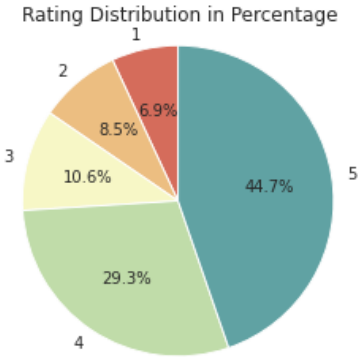


Fig 7: Distribution of rating in percentage

(B) Data Pre-processing

There are various preprocessing steps carried out to make the data ready to implement the topic modelling techniques.

Null value check: Before the model implementation, I have checked for the numm values in the dataset. This is done using the isnull() function, ant it is found that there are no null values in the chosen dataset.

```
df.isnull().sum()
```

```
Review    0
Rating    0
dtype: int64
```

Fig 8: Null value check

Stop word removal: One of the preprocessing techniques that is most frequently utilised is stop word removal. This is done to exclude words that appear frequently throughout all of the corpus's documents.
NLTK library is used in this research to remove stop words. If a term appears in the list of stop words provided by NLTK, it can be removed from a sentence by first dividing it into words.

```
import nltk
nltk.download('stopwords')

from nltk.corpus import stopwords
stop_words = stopwords.words('english')

def remove_stopwords(text):
    textArr = text.split(' ')
    rem_text = " ".join([i for i in textArr if i not in stop_words])
    return rem_text

# remove stopwords
df['Review']=df['Review'].apply(remove_stopwords)
```

Fig 9: Removal of stop words

Lemmatization: Lemmatization is used for the data normalisation process, which basically combines various forms of the same word to make it available as a single word for the analysis purpose.
Spacy library is used in this research to perform the lemmatization. Code written for the implementation is updated in the appendix.

```
[ ]  import en_core_web_sm
     nlp = spacy.load('en_core_web_sm')
```

```
[ ]  nlp = spacy.load('en_core_web_sm', disable=['parser', 'ner'])

     def lemmatization(texts,allowed_postags=['NOUN', 'ADJ']):
            output = []
            for sent in texts:
                  doc = nlp(sent)
                  output.append([token.lemma_ for token in doc if token.pos_ in allow
            return output
```

```
text_list=df['Review'].tolist()
print(text_list[1])
tokenized_reviews = lemmatization(text_list)
print(tokenized_reviews[1])
```

```
nothing special charge diamond member hilton decided chain shot 20th anniversary
['special', 'charge', 'diamond', 'member', 'chain', '20th', 'anniversary', 'seat
```

Fig 10: Lemmatization implementation

Removing space and short words: There were many spaces and short words with length less than 3 in the dataset. This is to be removed as it is not required for the topic modelling. This is done by creating a separate function and cleaning up all unwanted short words and spaces.

```
def clean_text(text ):
    delete_dict = {sp_character: '' for sp_character in string.punctuation}
    delete_dict[' '] = ' '
    table = str.maketrans(delete_dict)
    text1 = text.translate(table)
    #print('cleaned:'+text1)
    textArr= text1.split()
    text2 = ' '.join([w for w in textArr if ( not w.isdigit() and  ( not w.isdigit() and len(w)>3))])

    return text2.lower()
```

Fig 11: Function to remove space and short words

RESULTS

(A) LDA - Latent Dirichlet Allocation

LDA is implemented using the “LdaModel” function which is from the “genesis" library. After the data processing, a data dictionary is created, here the dictionary is called “Corpus”. For implementing

the LDA modelling, I have set the number of topics as 5. After creating 5 different topic groups, the result looks as below.

```
lda_model.print_topics()
```

```
[(0,
  '0.067*"beach" + 0.062*"resort" + 0.047*"pool" + 0.047*"good" + 0.026*"great" + 0.025*"water" + 0.023*"people" + 0.021*"nice" + 0.020*"restaurant" + 0.015*"drink"'),
 (1,
  '0.032*"food" + 0.031*"great" + 0.030*"beautiful" + 0.026*"staff" + 0.021*"time" + 0.021*"service" + 0.020*"wonderful" + 0.018*"fantastic" + 0.017*"restaurant" +
0.015*"trip"'),
 (2,
  '0.029*"hotel" + 0.020*"service" + 0.018*"food" + 0.018*"time" + 0.015*"room" + 0.014*"staff" + 0.012*"hour" + 0.010*"people" + 0.009*"guest" + 0.009*"thing"'),
 (3,
  '0.149*"hotel" + 0.052*"great" + 0.041*"location" + 0.040*"good" + 0.030*"staff" + 0.028*"room" + 0.022*"breakfast" + 0.020*"clean" + 0.018*"nice" +
0.018*"friendly"'),
 (4,
  '0.080*"room" + 0.020*"night" + 0.017*"hotel" + 0.016*"area" + 0.013*"nice" + 0.012*"small" + 0.012*"floor" + 0.011*"bathroom" + 0.009*"good" + 0.009*"breakfast"')]
```

Fig 12: Topics

After implementing the LDA topic modelling the result is plotted using the PyLDAvis library. Using this library we can picturize the result in a multidimensional scaling in an intertopic distance map. Most relevant terms for the topic with the data percentage will be displayed in the adjacent bar graph up on selecting any topic from the result set.

Result cluster set view (Intertopic distance map) is shown in figure 13. From this view it is clear that the LDA topic modelling technique was able to do the clustering for the topics 1, 2, 4 and 5. But topic 3 got mixed up with topic 1.

```
import pyLDAvis
import pyLDAvis.gensim_models

pyLDAvis.enable_notebook()
vis = pyLDAvis.gensim_models.prepare(lda_model, doc_term_matrix, dictionary)
vis
```

```
/usr/local/lib/python3.7/dist-packages/pyLDAvis/_prepare.py:247: FutureWarning: In a future version of pandas all arguments of DataFrame.drop except for the argument 'labels' will be keyword-only
  by='saliency', ascending=False).head(R).drop('saliency', 1)
```

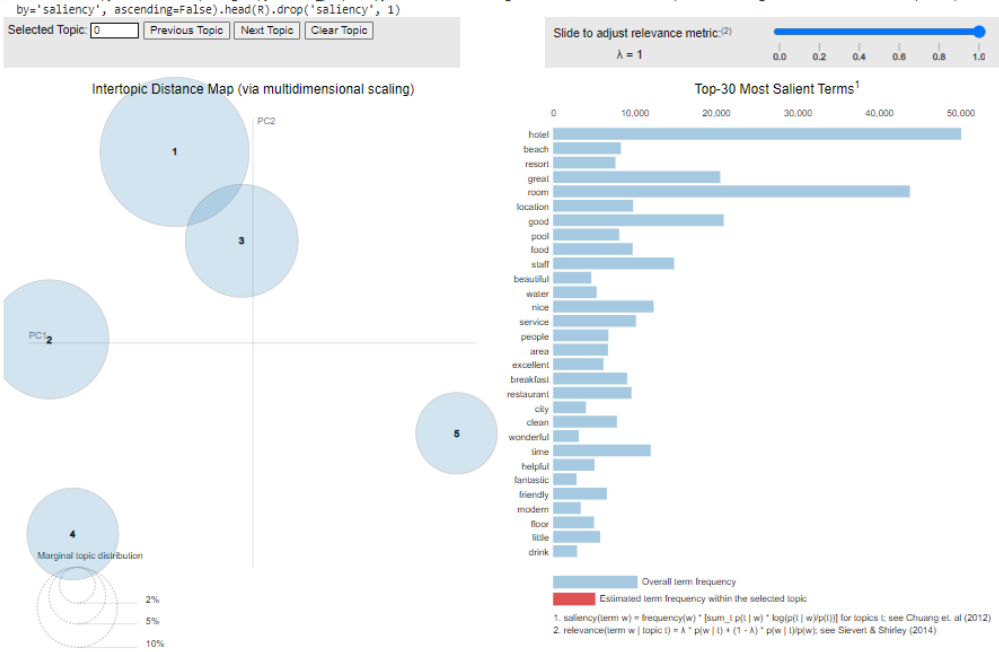


Figure 13: Clustering using PyLDAvis

(B) LSA - Latent Semantic Analysis

The next topic modelling technique used in this study is LSA. The preprocessing steps carried out for the LDA technique is used before implementing LSA modelling to make the data ready. Then I have written a function to find the top 15 words from the document list and their number of occurrences. It is then plotted using a bar graph, which is shown the figure 14.

I have created 8 topics for the LSA implementation and the result set is as below.

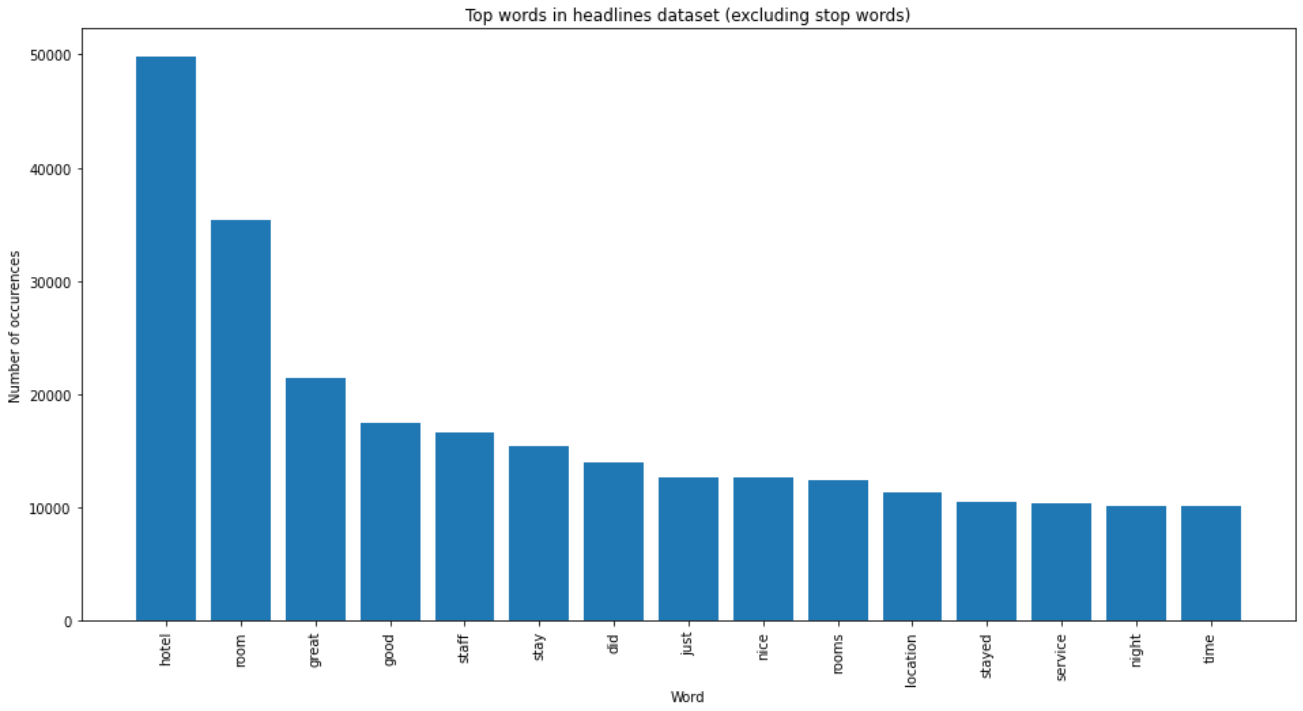


Figure 14: Top 15 words

I have created 8 topics for the LSA implementation and the result set is as below (figure 15).

```
for i in range(len(top_n_words_lsa)):
    print("Topic {}: ".format(i+1), top_n_words_lsa[i])
```

```
Topic 1:  [b'hotel', b'room', b'great', b'good', b'staff', b'stay', b'did', b'nice', b'ju
Topic 2:  [b'resort', b'beach', b'food', b'people', b'like', b'did', b'vacation', b'cours
Topic 3:  [b'room', b'bed', b'hotel', b'stay', b'night', b'did', b'small', b'desk', b'flo
Topic 4:  [b'_\xc3\xa3', b'room', b'food', b'hotel', b'_\xc3\xa2', b'it__\xc3\xa3', b'bea
Topic 5:  [b'great', b'location', b'staff', b'rooms', b'stay', b'clean', b'room', b'staye
Topic 6:  [b'good', b'hotel', b'location', b'breakfast', b'clean', b'room', b'rooms', b's
Topic 7:  [b'did', b'people', b'just', b'sleep', b'money', b'place', b'like', b'went', b'
Topic 8:  [b'stay', b'staff', b'stayed', b'excellent', b'place', b'friendly', b'location'
```

Figure 15: Top terms in each topic

From the result set it is clear that the top terms are getting repeated in each topic, which is different from what we have seen in the LDA model. The interfering of terms in each topic can be clearly visible in t-SNE clustering representation.

From the figure 16 (clustering result), it is evident that there is no separation between the topics. So when comparing with the LDA model, we can say that the result of LSA modelling is a bit dissatisfying.

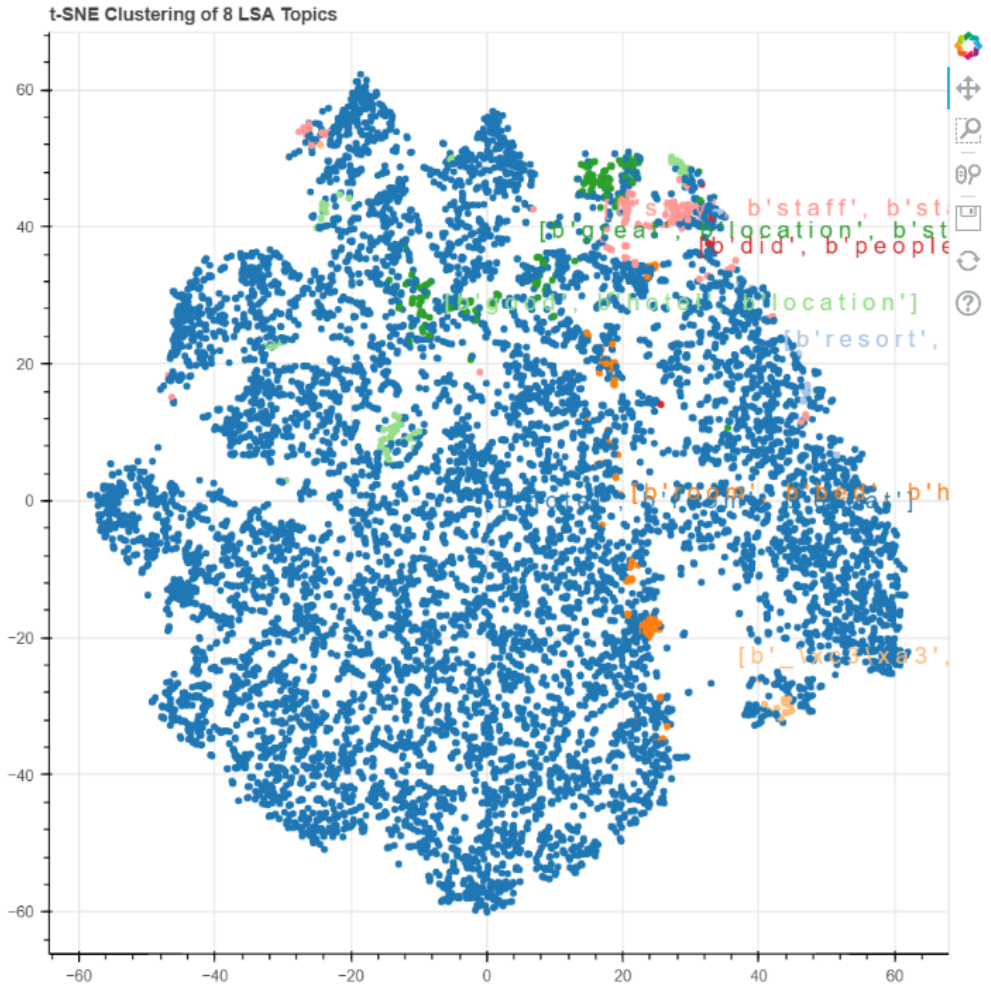


Figure 16: t-SNEclustering of LSA topics

(C) BERT (Bidirectional Encoder Representations from Transformers)

By using pre-trained transformer-based language models to embed each text, BERTopic expands on the concept of addressing topic modelling as a clustering task. BERTopic uses the cuml package to provide GPU support. The RAPIDS framework, which needs to be installed by conda or Docker, is the only way to install this package, though.

The next topic modelling technique used in this study is BERT. The preprocessing steps carried out for the LDA and LSA techniques are used before implementing BERT modelling to make the data ready.

I have created 11 topics to evaluate the BERT topic modelling and came to a conclusion that instead of having the sentiment analysis based on the review provided, the BERT modelling clusters the review based on the location, which is represented in the figure 17. And the accuracy we are getting for this model is 83.7 Which is relatively high accuracy.

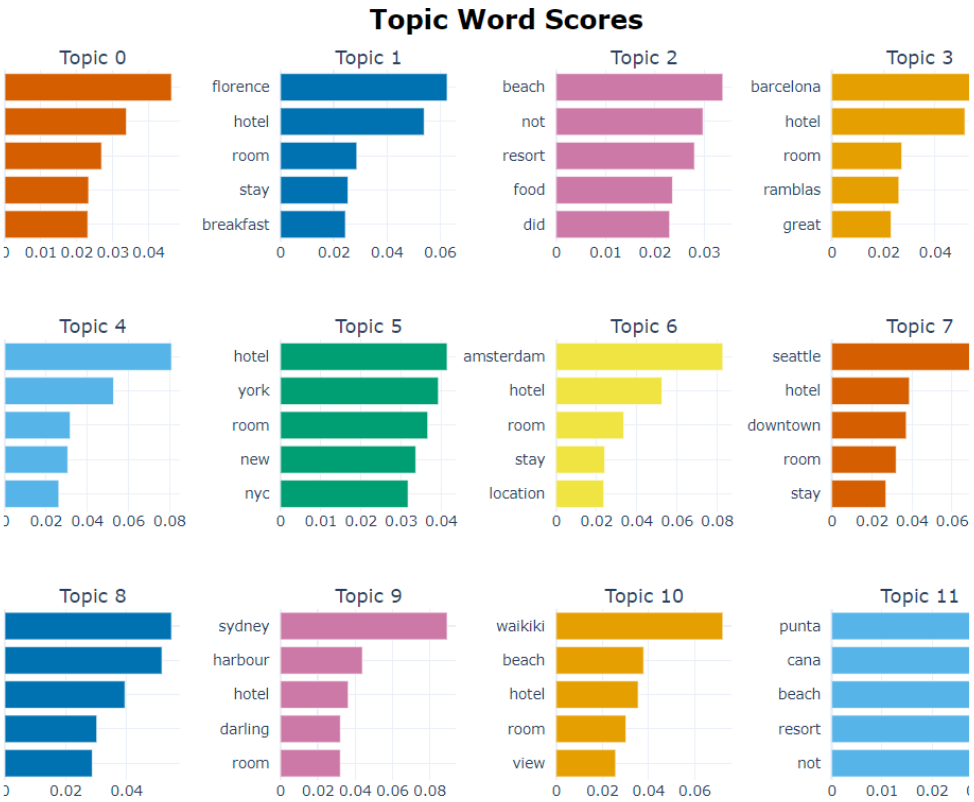


Figure 17: BERTopic clustering

CONCLUSION

In this paper, we have used three different topic modelling techniques. Among these, 3 based on the topic cluster creation it is evident that the LDA model is comparatively performing better than the LSA model. In the LSA model during the topic clustering, all the clusters were intermixed and there was no proper differentiation among them. I have also done research on the BERT topic, where it was concluded that instead of sentimental clustering, BERT model have done the clustering based on the location of the hotel.

Based on enough evidence after the study, I came to a conclusion that the best performing model for the hotel review sentiment analysis is the LDA model.

ETHICAL CONSIDERATION

The ethical problems which is there in the review analysis dataset is similar to the ones in any data driven project. But as the collected review is publicly made available by the owner company there are no complications involved in utilising the hotel review dataset.

Task 2 : Evolutionary and Fuzzy Systems

Part 1 – Design and Implementation of the FLC

Introduction

A fuzzy control system is an algorithm-based control system. Fuzzy logic allows us to create complicated behaviour logic using simple rules.
I have used the MATLAB fuzzy logic toolbox package used in this research to implement an intelligent assistive care environment. There are 7 inputs and 6 different outputs created to design fuzzy logic. It is implemented in an acer 11th Gen Intel(R) Core(TM) i5 using the windows 10 operating system.

We have to consider multiple factors while implementing the intelligent flat for the disabled people. Since the users of the flat need special care sensors are installed for various purposes such as to detect the movement of persons, to auto-adjust the temperature and light and the sensors to detect the medical attention.

The main component of this study is the input and the functions. Each input individually or combined with other inputs generates the output.

Input and Its Membership Functions:

There are 7 different inputs created for design; those are temperature, humidity, outside light, movements, body temperature, heart rate and CO2 level in the air. An associated membership function is maintained for each input based on the requirement, and an intelligent flat can be created by using these inputs in the design. Details of input and its membership methods are as given below.

(i) Temperature:

One of the main inputs in the design of the intelligent flat is temperature. Here it refers to the temperature outside the room, based on which the temperature inside the flat will be auto adjusted. This is one of the main parameters which is used to regulate the air conditioner and the room heater. Temperature is divided into 5 values; Very low, low, medium, high and very high. If the temperature

is going above or below the normal level would make the people in the flat uncomfortable. So the inhouse warmth is maintained based on the outside temperature.

Figure 1 showcases the membership function designed and table 1 explains the range selected.

There are 2 types of functions defined here: trapezoid and triangle. Where trapezoid is used for very high and very low and triangle is used for high, medium and low.

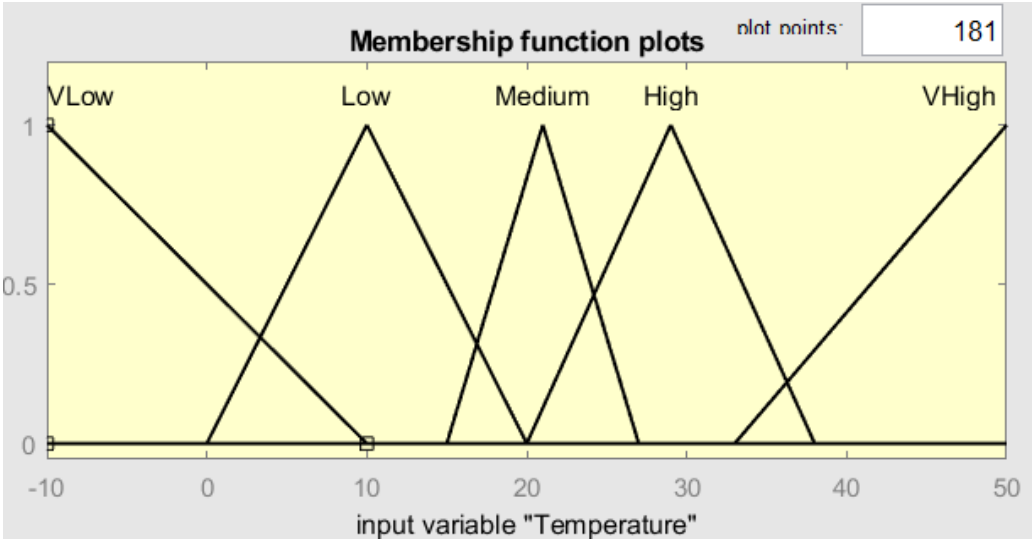


Fig 1: Membership function for Temperature

Input variables	Range	MF Type
Very Low	-10°C to 10°C	Trapezoid
Low	0°C to 20°C	Triangle
Medium	15°C to 27°C	Triangle
High	20°C to 38°C	Triangle
Very High	33°C to 50°C	Trapezoid

Table 1: Temperature range

(ii) Humidity:

Humidity also plays an important role in adjusting the air conditioner and room heater in the flat, as the humidity goes higher, people will feel hotter than the actual temperature. There are 5 values set like the temperature input; Very low, low, medium, high and very high.

Figure 2 showcases the membership function designed for humidity and table 2 explains the range selected.

Input variables	Range	MF Type
Very Low	0% to 25%	Trapezoid
Low	20% to 45%	Triangle
Medium	40% to 65%	Triangle
High	60% to 80%	Triangle
Very High	70% to 100%	Trapezoid

Table 2: Humidity range

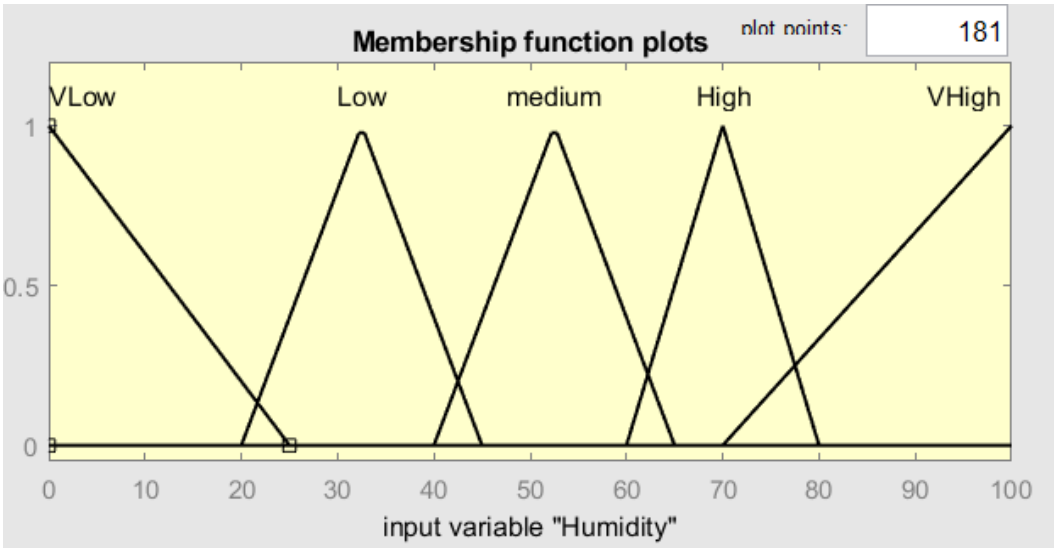


Fig 2: Membership function for Humidity

There are 2 types of functions defined here: trapezoid and triangle. Where trapezoid is used for very high and very low and triangle is used for high, medium and low.

(iii) Outside Light:

There are two automations controlled by this parameter, and they are window blind and room light. If there is enough sunlight then there is no need to illuminate the bulb in the room and the blinds should be open inorder for the light to reach the room. There are 3 different values defined for this input- Dark, normal and bright. 2 types of functions are defined - trapezoid and triangle. Trapezoid is used for dark and bright light, triangle is defined for normal light.

Figure 3 showcases the membership function designed for outside light and table 3 explains the range selected.

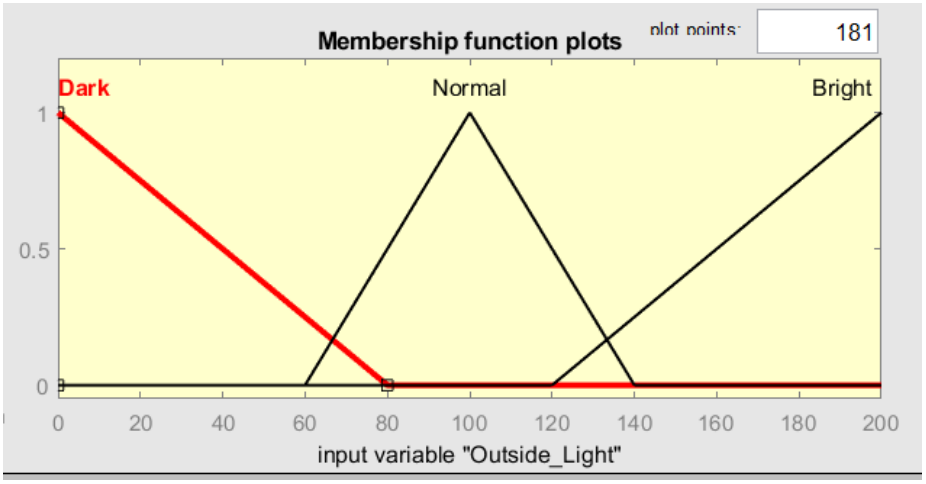


Fig 3: Membership function for Outside Light

Input variables	Range	MF Type
Dark	0 to 80	Trapezoid
Normal	60 to 140	Triangle
Bright	120 to 200	Trapezoid

Table 3: Outside Light range

(iv) Movement:

Any security system needs motion sensors as a crucial part. A sensor will initiate an alarm notification to the caretaker or to the medical department when it notices any unusual movement. Unusual movements meant here is any movement which violates the pattern. There will be a movement pattern saved in the system like the normal walking or doing exercises. Any movement which the system detects other than the usual movement will send an alarm. There are 3 different values defined for this input- No movement, normal and unusual. 2 types of functions are defined - trapezoid and triangle. Trapezoid is used for none and unusual, triangle is defined for normal movements.

Figure 4 showcases the membership function designed for movement and table 4 explains the range selected.

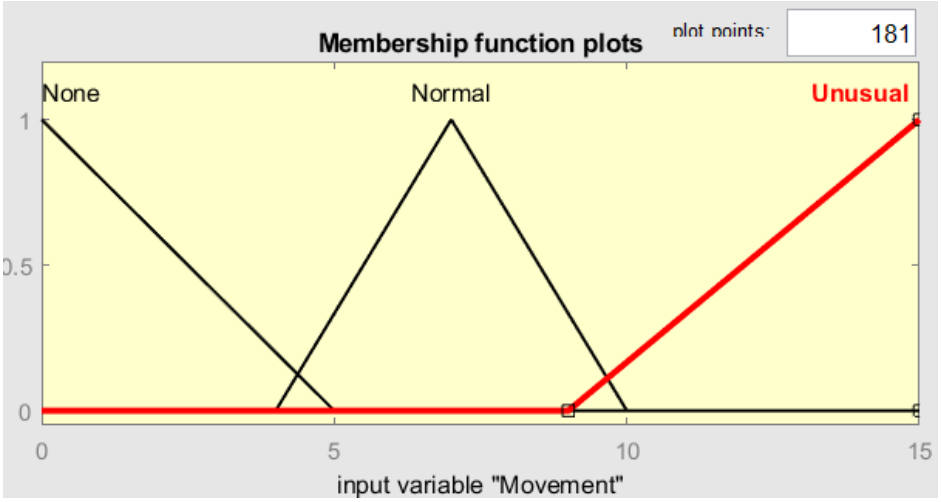


Fig 4: Membership function for Movement

Input variables	Range	MF Type
No Movement	0 to 5	Trapezoid
Usual Movements	4 to 10	Triangle
Unusual Movements	9 to 15	Trapezoid

Table 4: Movement range

(v) Body Temperature:

Body temperature is an important parameter here as the occupants are disabled or ill. Monitoring the body temperature and sending a corporal alarm will help the caretaker or the medical department to assess the users medical condition and to give good care. There are 3 different values defined for this input- Low, Normal and High. 2 types of functions are defined - trapezoid and triangle. Trapezoid is used for Low and High, triangle is defined for normal body temperature.

Figure 5 showcases the membership function designed for body temperature and table 5 explains the range selected.

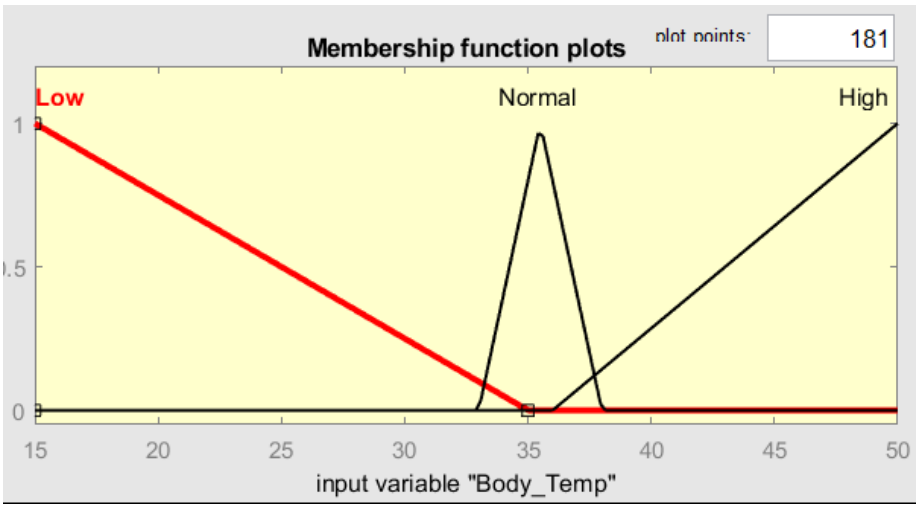


Fig 5: Membership function for Body temperature

Input variables	Range	MF Type
Low	15 to 35	Trapezoid
Normal	33 to 38	Triangle
High	36 to 50	Trapezoid

Table 5: Body Temperature range

(vi) Heart Rate:

Heart rate monitoring is an important part in automating an assistive care environment. As the occupants in this flat require special care, automatic detection of heart rate and alarming the medical crew if there is any variation will take the caretakers to give special attention and care. There are 3 different values set for this input- Low, Normal and High, range for this value is set based on the values medical practitioners set for the normal and abnormal heart rates.

Figure 6 showcases the membership function designed for heart rate and table 6 explains the range selected.

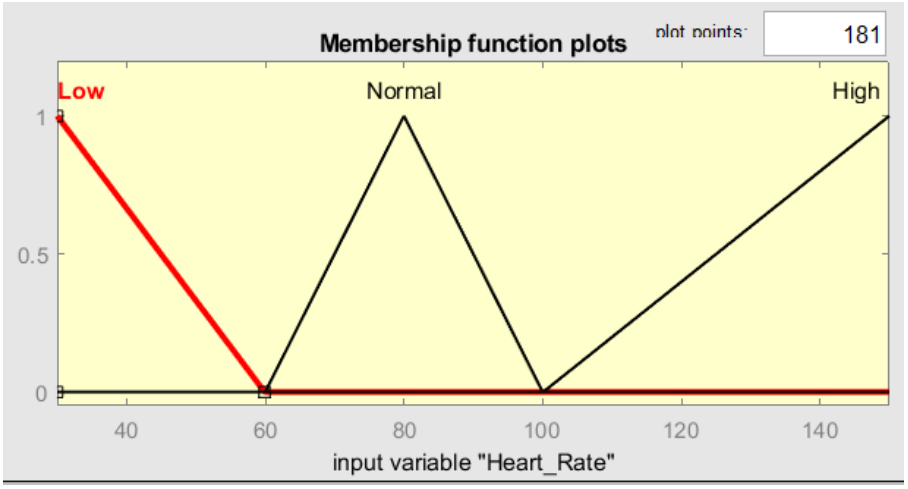


Fig 6: Membership function for Heart rate

Input variables	Range	MF Type
Low	30 to 59.9	Trapezoid
Normal	60 to 100	Triangle
High	100.1 to 150	Trapezoid

Table 6: Body Heart rate range

(vii) CO2:

The CO2 sensor is designed to measure the amount of carbon dioxide in the air. It is important to have a check on the CO2 level as high concentration of the same is harmful for humans and will affect multiple internal organs. Sensor will monitor the CO2 level and will send out an alarm if the concentration goes beyond the normal level. There are 3 different values set for this input- Low, Normal and High, range for this value is set based on the allowed percentage of abnormal CO2 level in the atmosphere.

Figure 6 showcases the membership function designed for CO2 sensor and table 6 explains the range selected.

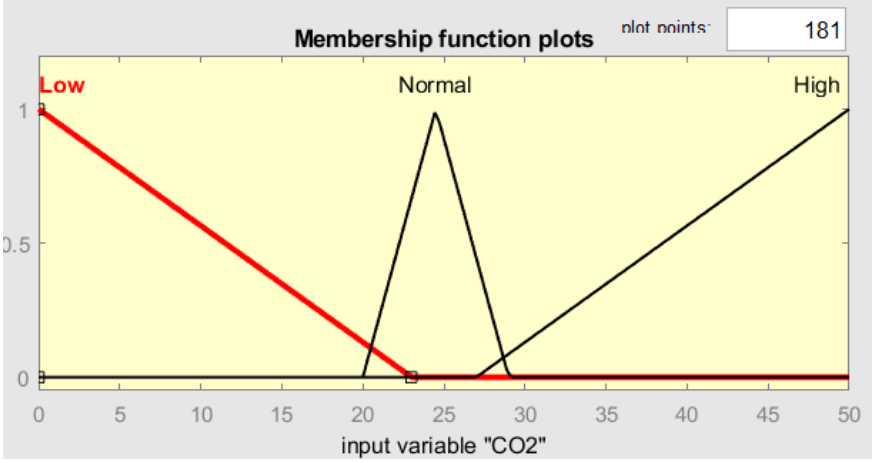


Fig 7: Membership function for CO2

Input variables	Range	MF Type
Low	0 to 23	Trapezoid
Normal	20 to 29	Triangle
High	27 to 50	Trapezoid

Table 7: CO2 concentration range

Fuzzy Inference System:

The fuzzy inference system is the central component of the fuzzy logic system which primarily focuses on the decision making. It uses conditional statements like If, then and boolean logical operators like AND, OT, NOT for creating rules based on a single or combination of multiple inputs for the decision making. No matter how crisp or fuzzy the input is to FIS, the output is always a fuzzy set. When it is used as a controller, fuzzy output is required. For the purpose of converting fuzzy variables into crisp variables, a defuzzification unit would be present with FIS.
There are two different types of inference systems available.

1. Mamdani inference system
2. Sugeno inference system.

In this paper, I have used the “mamdani” inference system for the study. MATLAB fuzzy control toolbox was used to implement the same.

The Mamdani fuzzy inference system implemented for this study is as shown in the figure 8.

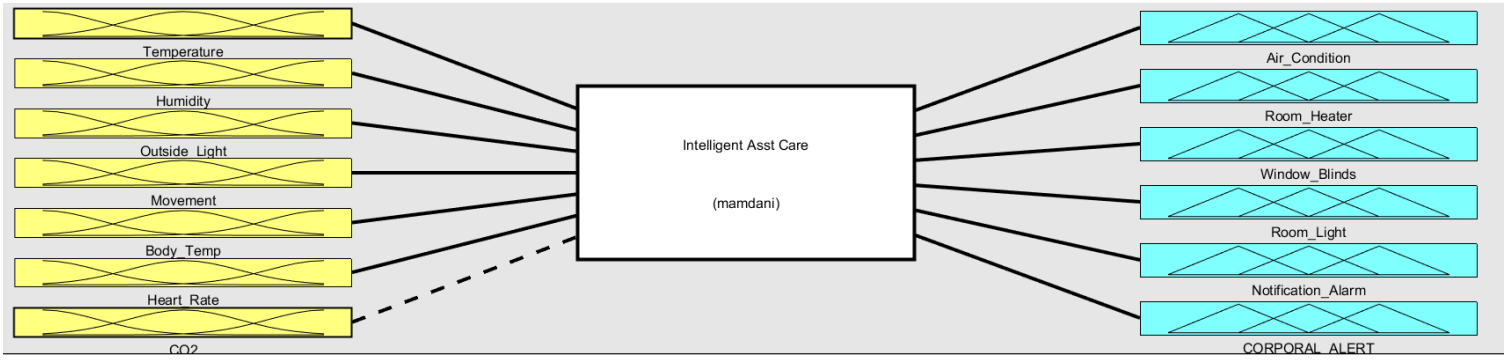


Fig 8: Mamdani fuzzy inference system

Output and membership function:

There are 6 outputs in the designed FLC, those are Air condition, room heater, window blinds, room light, notification alarm, and corporal alert.

(i) Air conditioner:

Air conditioner is used to adjust the temperature in the room when there is hot weather outside. This output is created by considering the temperature and humidity in the iput. There are 5 member functions defined and those are Very low, low, medium, high and very high. Out of these functions extreme conditions are represented in trapezoidal and other 3 functions are represented in triangle.

Figure 9 showcases the membership function designed for the air conditioner and table 9 indicates the range.

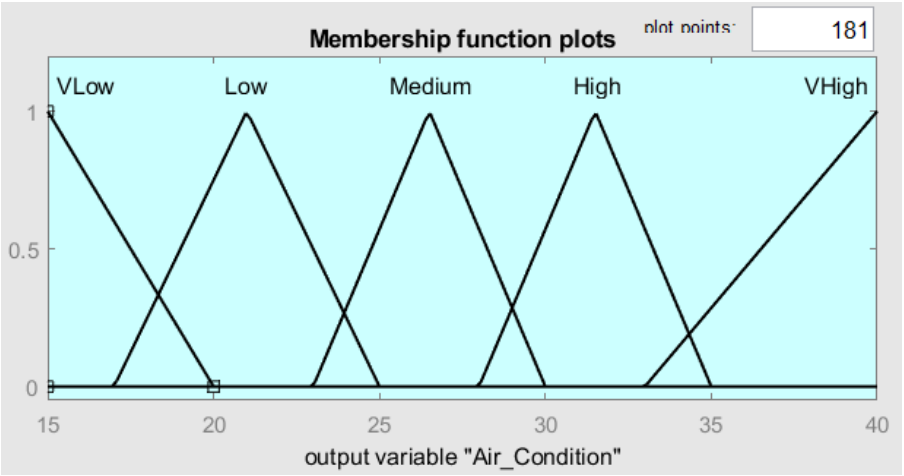


Fig 9: Membership function for air conditioner

Output variables	Range	MF Type
Very Low	15°C to 20°C	Trapezoid
Low	17°C to 25°C	Triangle
Medium	23°C to 30°C	Triangle
High	28°C to 35°C	Triangle
Very High	33°C to 40°C	Trapezoid

Table 9: Air conditioner range

(ii) Room Heater:

When it's cold outside, the output of a room heater is employed to keep the space warm. This output contains total 5 functions out of which 2 are trapezoidal (very high and very low), and other 3 are triangle (high, low and medium)

Figure 10 showcases the membership function for the room heater and table 10 indicates the range.

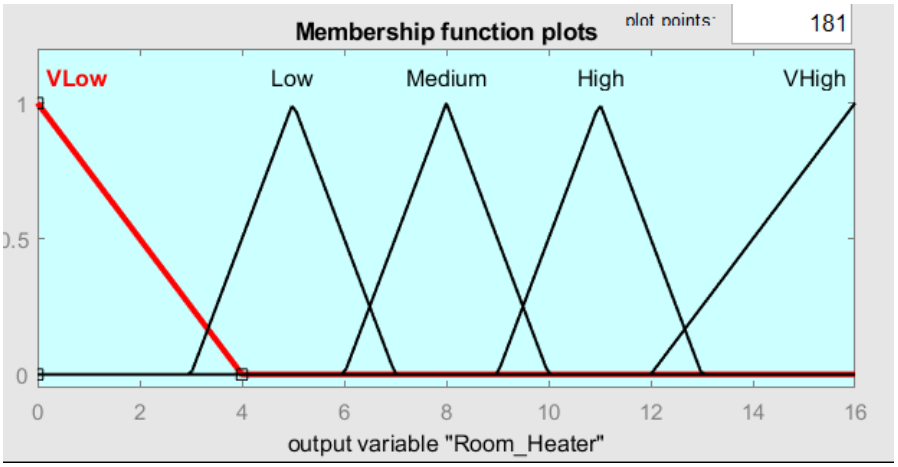


Fig 10: Membership function for room heater

Output variables	Range	MF Type
Very Low	0 to 4	Trapezoid
Low	3 to 7	Triangle
Medium	6 to 10	Triangle
High	9 to 13	Triangle
Very High	12 to 16	Trapezoid

Table 10: room heater range

(iii) Window blinds:

Window blinds work based on the outside light and this output is designed to control the window blinds. This is used for adjusting the natural light inside the flat. There are 3 member functions designed here and among these 2 are trapezoidal (fully closed and fully open) and one is triangle (half open).

Figure 11 showcases the membership function for the window blinds and table 11 indicates the range.

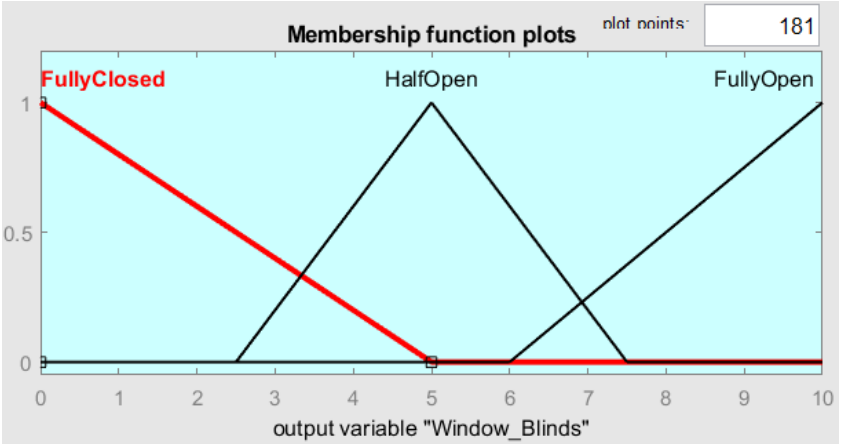


Fig 11: Membership function for window blinds

Output variables	Range	MF Type
Fully Closed	0 to 5	Trapezoid
Half Open	2.5 to 7.5	Triangle
Fully Open	6 to 10	Trapezoid

Table 11: window blinds range

(iv) Room light:

Room light automation is used for illuminating the light when it becomes dark in the room and vice versa. There are 3 member functions designed here and among these 2 are trapezoidal (Dark and bright light) and one is triangle (dim light).

Figure 12 showcases the membership function for the room light and table 12 indicates the range.

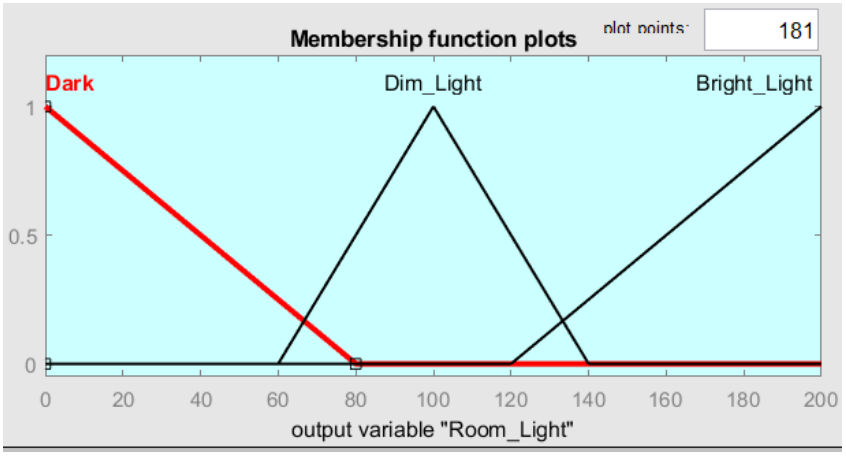


Fig 12: Membership function for room light

Output variables	Range	MF Type
Dark	0 to 80	Trapezoid
Dim Light	60 to 140	Triangle
Bright Light	120 to 200	Trapezoid

Table 12: room light range

(v) Notification alarm:

Notification alarm is used for alerting the caretaker for any unusual activity which needs attention from the crew. There are 3 member functions designed here and among these 2 are trapezoidal (No alert and high alert) and one is triangle (Low alert).

Figure 13 showcases the membership function for the notification alarm and table 13 indicates the range.

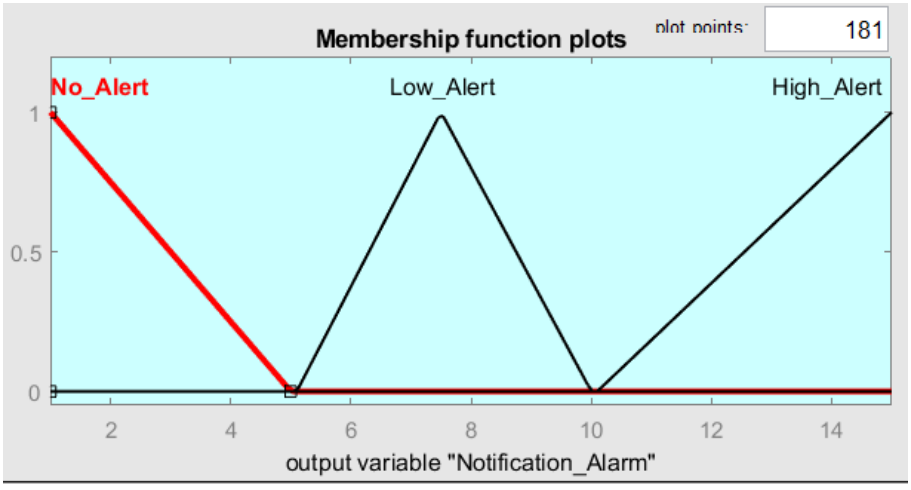


Fig 13: Membership function for Notification alarm

Output variables	Range	MF Type
No Alert	1 to 5	Trapezoid
Low Alert	5.1 to 10	Triangle
High Alert	10.1 to 15	Trapezoid

Table 12: Notification alarm range

(vi) Corporal Alert:

Corporal alert is used for alerting the caretaker or medical crew for any abnormal medical condition that occurs for the occupants in the flat. There are 3 member functions designed here and among these 2 are trapezoidal (No alert and high alert) and one is triangle (Low alert).
Figure 14 showcases the membership function for the corporal alarm and table 13 indicates the range.

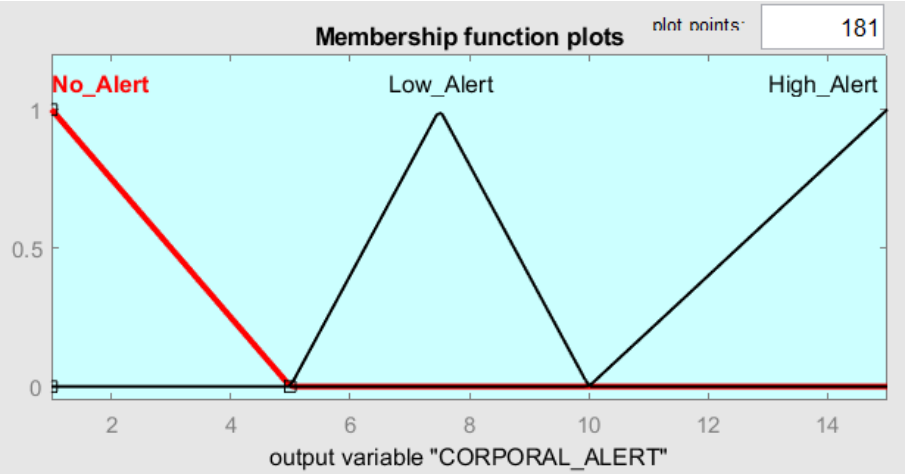


Fig 14: Membership function for corporal alert

Output variables	Range	MF Type
No Alert	1 to 5	Trapezoid
Low Alert	5.1 to 10	Triangle
High Alert	10.1 to 15	Trapezoid

Table 14: corporal alert range

Fuzzy Rules:

These are conditional statements which determine the output variable's value depending on the values of its input variables. These rules are written using "if.. Then" conditional statements and using the logical operators such as "AND, OR, NOT". I have designed a total of 43 rules in this FLC, which is as follows.

(i) Rule for Air condition and room heater

Humidity / Temperature	VLow	Low	Medium	High	VHigh
VLow	VHigh	VHigh	High	High	High
Low	High	High	High	Medium	Medium
Medium	Medium	Medium	Medium	Medium	L
High	Low	Low	Low	Low	Low
VHigh	VLow	VLow	VLow	VLow	VLow

(ii) Rule for Window Blind and Room Light

Outside_Light	Dark	Normal	Bright
Window Blind	Fully Closed	Half Open	Fully open

Outside_Light	Dark	Normal	Bright
Room Light	Bright Light	Half Open	Fully open

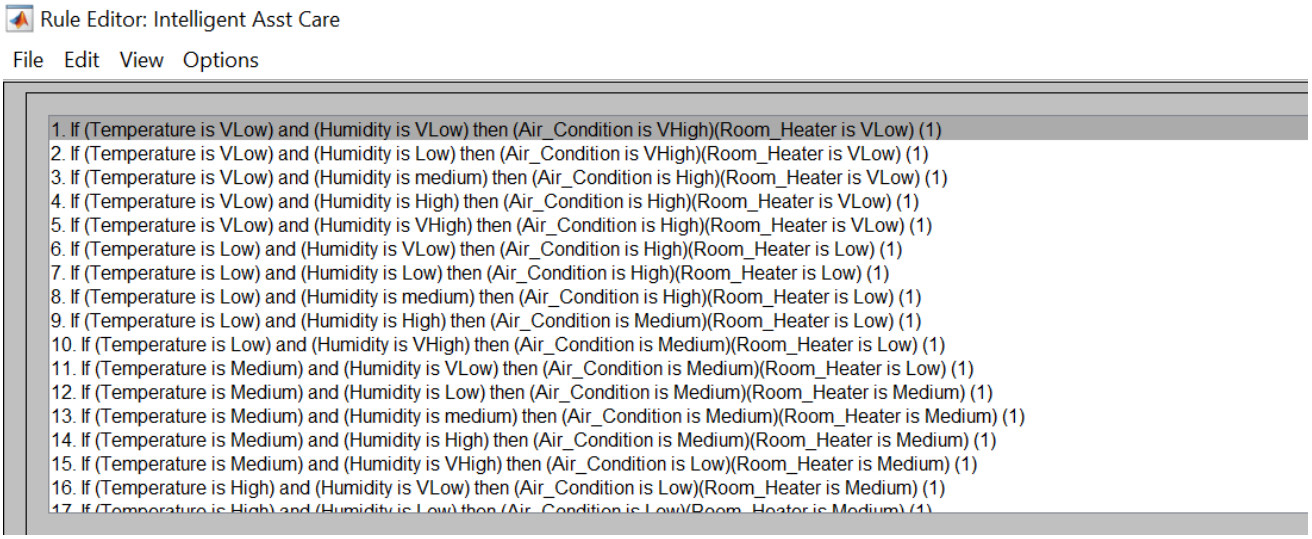
(iii) Rule for Unusual movement alert

Movement	None	Usual	Unusual
Notification	No Alert	Low Alert	High Alert

(iv) Rule for Unusual corporal alert

Body Temp / Heart Rate	**Low**	**Normal**	**High**
Low	High Alert	High Alert	High Alert
Normal	High Alert	No Alert	High Alert
High	High Alert	High Alert	High Alert

Rules written using the conditional and boolean operator is as follows:

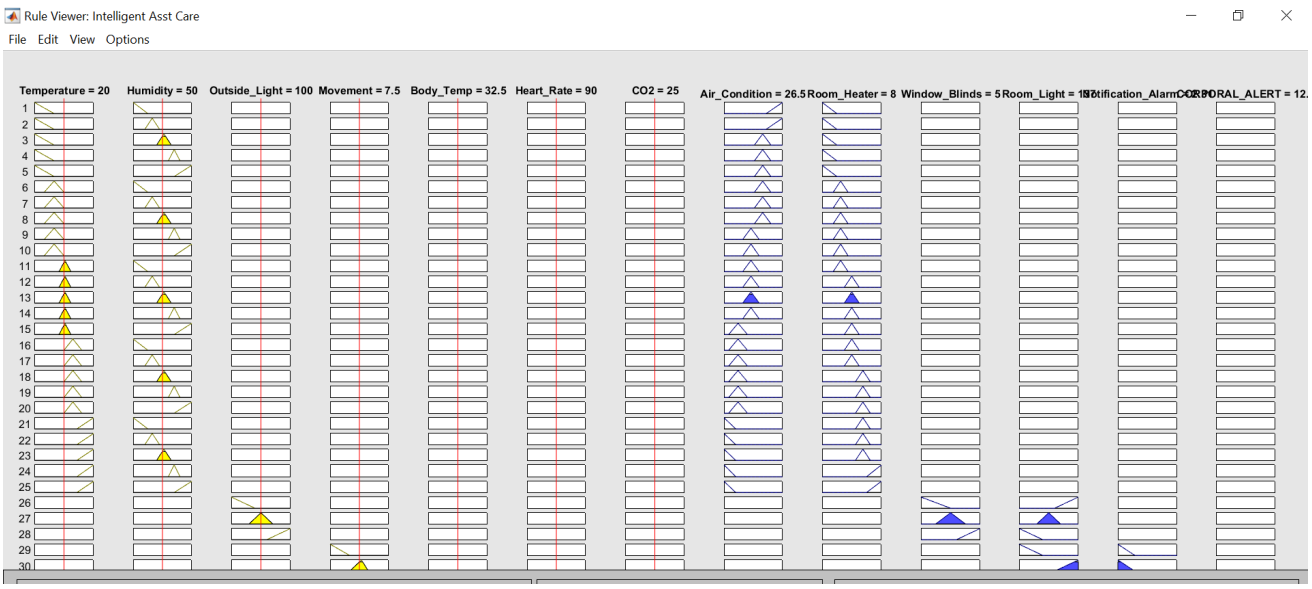


Aggregation:

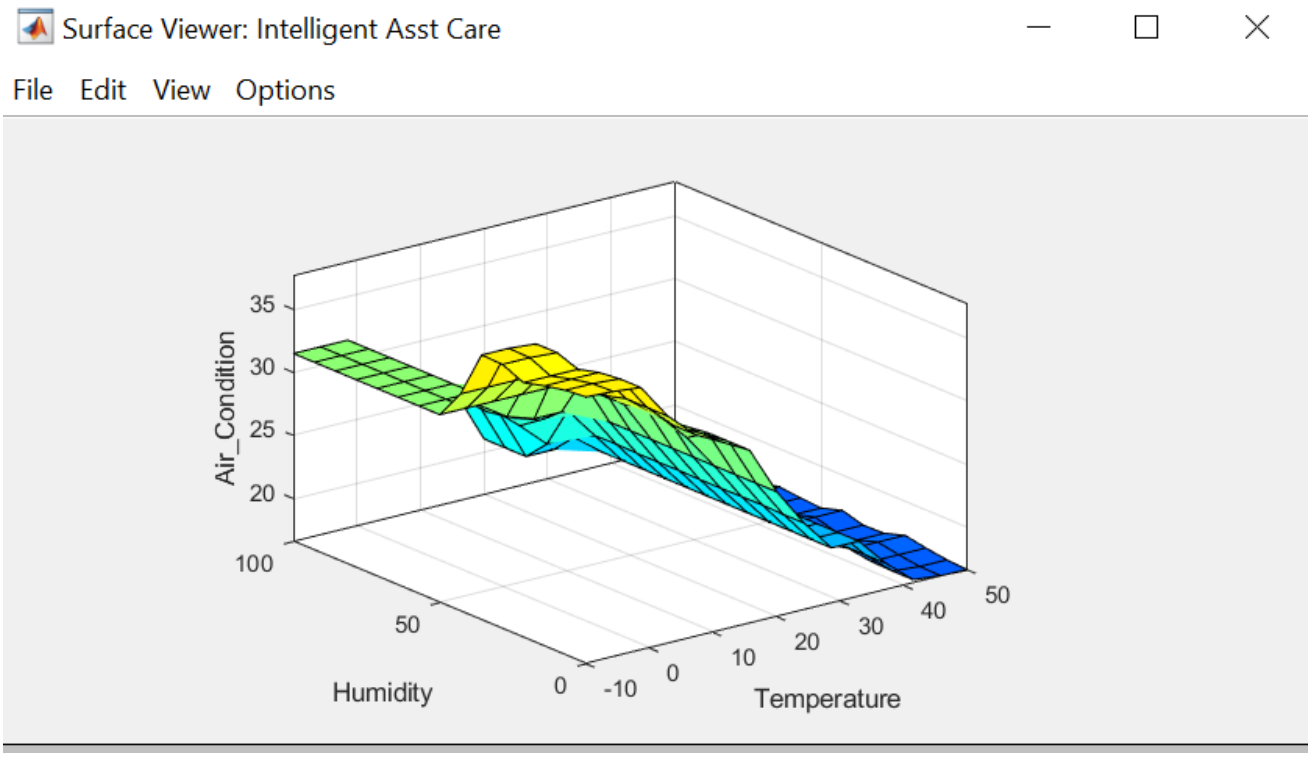
The act of consolidating the fuzzy sets that represent each rule's outputs into a single fuzzy set is known as aggregation.

Defuzzification:

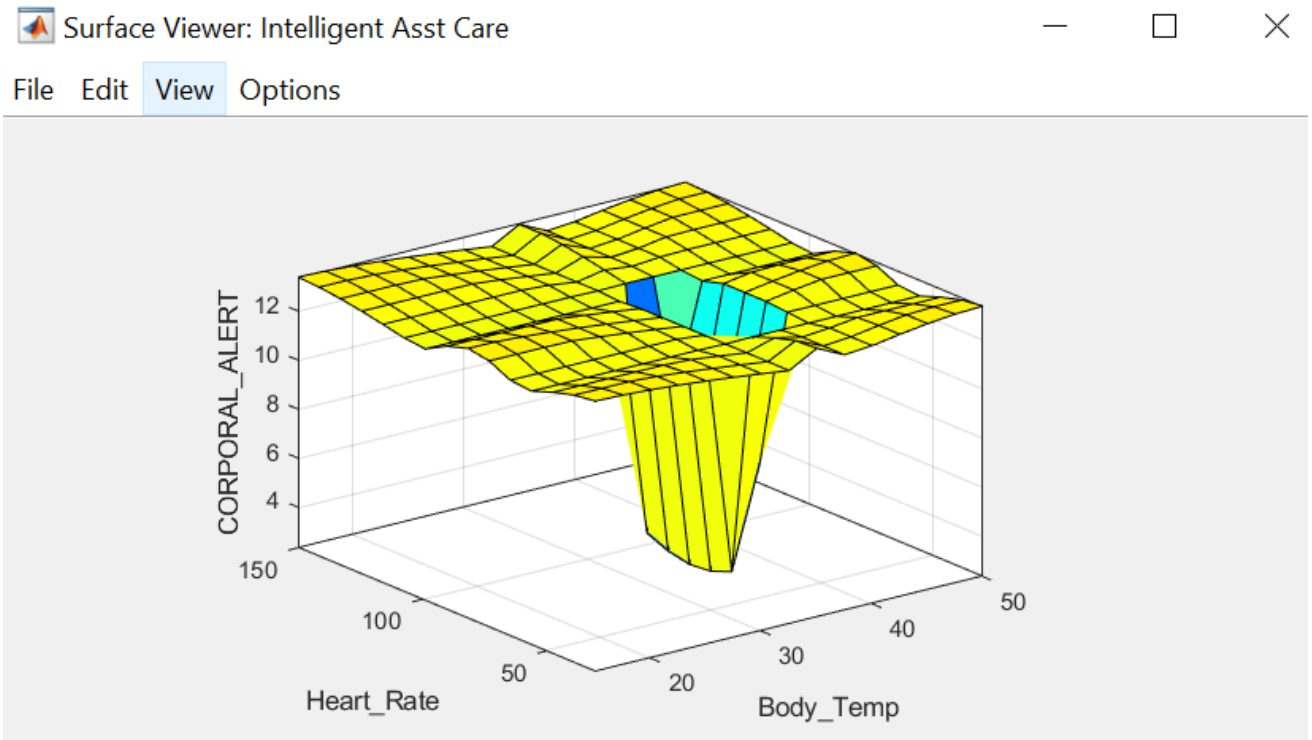
The method of defuzzification involves taking a single number from the output of the combined fuzzy set. It is used to convert the findings of fuzzy inference into a clear output. In other words, a decision-making algorithm that chooses the optimal crisp value based on a fuzzy set is known as defuzzification. I have used the “Centroid” method in this study for defuzzification. After this process written rules can be viewed in a single window in the following format.



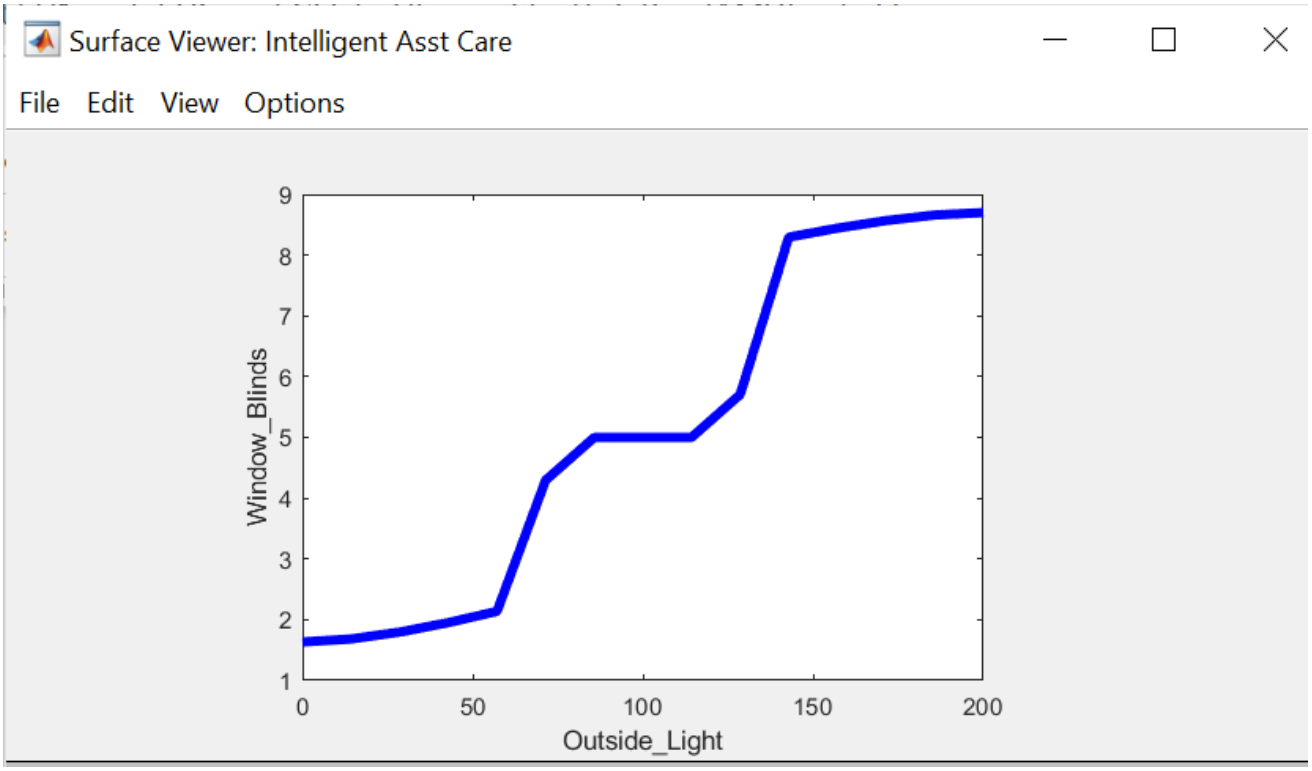
Below shows the surface view of air condition (input vs output)



Below shows the surface view of corporal alert (input vs output)



Below shows the surface view of window blind (input vs output)

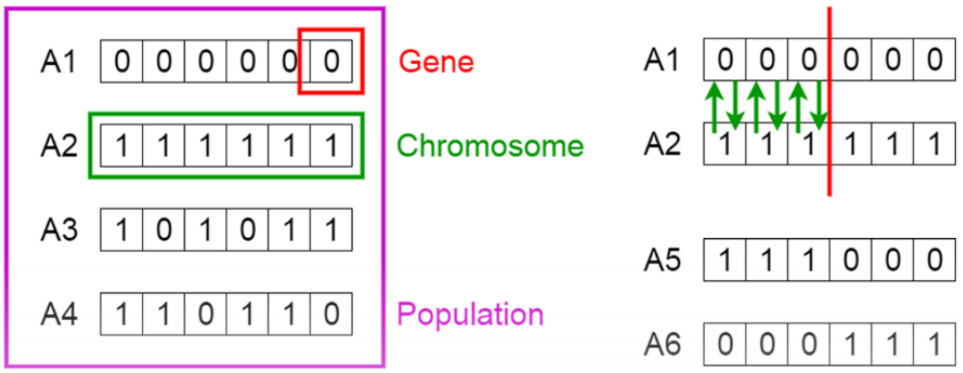


Conclusion:

In this study MATLAB fuzzy logic toolbox package was used to implement an intelligent assistive care environment. There are 7 inputs and 6 different outputs created to design fuzzy logic and 43 rules are created using the conditional and logical operators. With this design a fully automated intelligent flat can be implemented. Correct input and output function design plays a vital role here as the fuzzy system is used to address real world issues.
After creating the fuzzy system, the input and associated output can be tested and further optimised using the available algorithms.

PART 2 : Optimisation of FLC created in the first part:

The genetic algorithm, which is based on natural selection, is the mechanism that propels biological evolution and a technique for resolving both limited and unconstrained optimization issues. Genetic algorithm was developed after the inspiration from Charles Darwin's natural evolution theory. The algorithm makes use of bitstrings as analogues of a genetic representation, fitness as function evaluations, genetic recombination as bitstring crossovers, and mutation as mutation (flipping bits). Natural selection and genetics-based search algorithms are known as genetic algorithms.
Genetic algorithm has 5 steps those are:



1. Initial population : In this step, the initial population is selected. Here the problems are considered individually and generally known as genes, and the group of genes are known as a chromosome. To optimise our FLC process, we can consider each input and output as individual genes. To implement this each gene (FLC input or output) should be converted to a binary value.
2. Fitness function : Fitness of the genes selected in the initial process will be evaluated in this step. Based on the fitness result, each gene will be allocated an accurate value and will be finalised the one to be considered for the reproduction .
3. Selection : After considering the scores assigned, the best parent genes will be selected in this step. There are various methods available to select the best fit genes, and one of them is rank selection method.
4. Crossover : This is the crucial step in the genetic algorithm where the parent genes are randomly selected for the crossover.
5. Mutation : In this step the mutation will be done, and there are multiple methods available, and can be selected based on the needs.

Part 3: Comparing different optimisation techniques on the CEC2005 functions

There are multiple functions available, among which I chose **Rastrigin function** and **Ackley function**

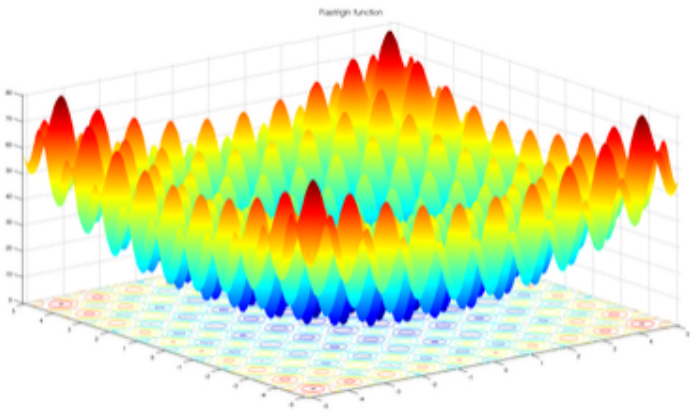
The Rastrigin function is a non-convex function that is used as a performance test issue for optimization techniques in mathematics. It is an illustration of a non-linear multimodal function in general. Due to the function's wide search space and numerous local minima, determining its minimum is a challenging job.

This function is defined using the following formula

$$f(\mathbf{x}) = An + \sum_{i=1}^{n} \left[x_i^2 - A\cos(2\pi x_i)\right]$$

where $A = 10$ and $x_i \in [-5.12, 5.12]$. It has a global minimum at $\mathbf{x} = \mathbf{0}$ where $f(\mathbf{x}) = 0$

The function can be represented as follows in a 3D graph:

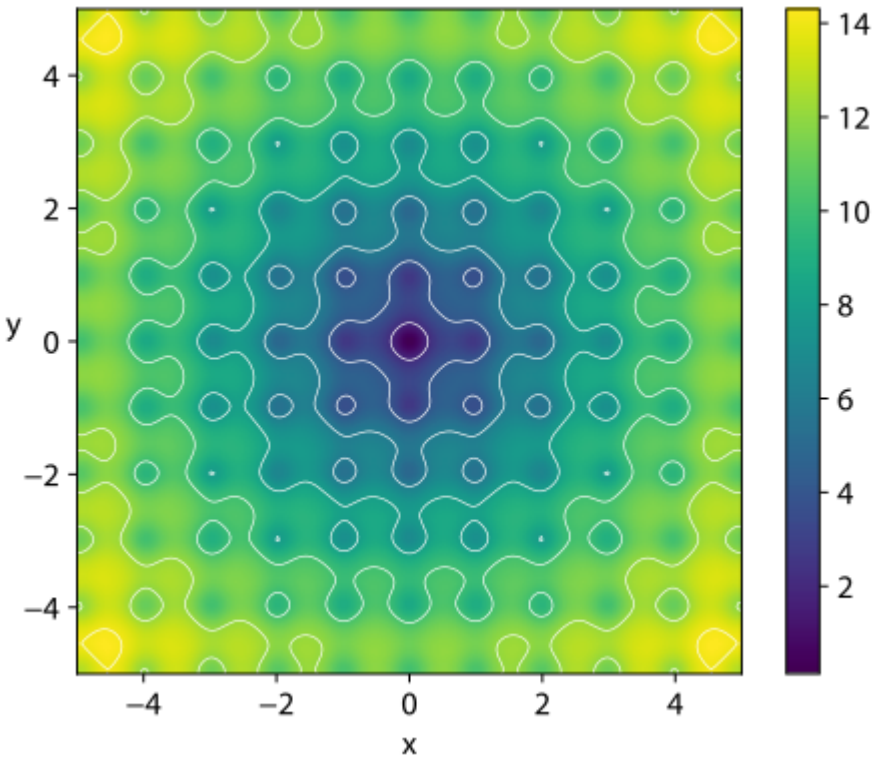


The second function which is the Ackley function is a non-convex function that is used as a performance test issue for optimization techniques in mathematics. Two dimension of this function is defined as:

$$f(x, y) = -20\exp\left[-0.2\sqrt{0.5(x^2 + y^2)}\right] - \exp[0.5\left(\cos 2\pi x + \cos 2\pi y\right)] + e + 20$$

Where the universal opium point f(0,0) = 0

The function can be represented as follows in a 2 dimensional graph:



APPENDIX

Code written for the implementation of topic modelling is publically available in the github repository. Link to access the code is :

https://github.com/priyankaharidasnk/Modelling

Link to the dataset is : https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

REFERENCE

1. HDS. 2022. *Topic Modeling with LDA Explained: Applications and How It Works - HDS*. [online] Available at: <https://highdemandskills.com/topic-modeling-intuitive/>
2. 2022. [online] Available at: <https://monkeylearn.com/blog/introduction-to-topic-modeling/#:~:text=Topic%20modeling%20is%20an%20unsupervised,characterize%20a%20set%20of%20documents>
3. Medium. 2022. *Topic Modelling using LDA*. [online] Available at: <https://medium.com/analytics-vidhya/topic-modelling-using-lda-aa11ec9bec13>
4. Zvornicanin, E., 2022. *Topic Modeling and Latent Dirichlet Allocation (LDA) | DataScience+*. [online] Datascienceplus.com. Available at: <https://datascienceplus.com/topic-modeling-and-latent-dirichlet-allocation-lda/>

5. Analytics Vidhya. 2022. *Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/>
6. Medium. 2022. *Topic Modeling and Latent Dirichlet Allocation (LDA) in Python*. [online] Available at: <https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24>
7. 2022. [online] Available at: <https://monkeylearn.com/blog/introduction-to-topic-modeling/>
8. Ieeexplore.ieee.org. 2022. *A sentiment analysis model for hotel reviews based on supervised learning*. [online] Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6016866>
9. Ieeexplore.ieee.org. 2022. *IEEE Xplore Full-Text PDF:*. [online] Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9210480>
10. Ieeexplore.ieee.org. 2022. *Sentiment Analysis of Visitor Reviews on Hotel In West Sumatera*. [online] Available at: <https://ieeexplore.ieee.org/document/9631859>
11. Medium. 2022. *Topic Modeling with LSA, PSLA, LDA & lda2Vec*. [online] Available at: <https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05>
12. Analytics Vidhya. 2022. *Topic Modelling using LSA | Guide to Master NLP (Part 16)*. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/06/part-16-step-by-step-guide-to-master-nlp-topic-modelling-using-lsa/>
13. ACM Other conferences. 2022. *Topic Modeling: Comparison of LSA and LDA on Scientific Publications | 2021 4th International Conference on Data Storage and Data Engineering*. [online] Available at: <https://dl.acm.org/doi/abs/10.1145/3456146.3456156>
14. Nlp.stanford.edu. 2022. *Stemming and lemmatization*. [online] Available at: <https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html>
15. Kaggle.com. 2022. *hotel reviews - topic modeling*. [online] Available at: <https://www.kaggle.com/code/sukhadadharangaonkar/hotel-reviews-topic-modeling>
16. https://www.kaggle.com/code/anitaclment/topic-modelling-using-bertopic
17. https://en.wikipedia.org/wiki/Rastrigin_function
18. https://en.wikipedia.org/wiki/Ackley_function
19. https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3
20. https://www.kindsonthegenius.com/basics-of-genetic-algorithm-ga-explained-in-simple-terms/