CHAPTER 4

INITIAL FINDINGS

4.1 Introduction

The conceptual framework is the framework for carrying out the detailed sentiment analysis for hotel reviews. This starts with the collection of data, then by data preprocessing, feature extraction and finally the implementation of a sentiment algorithm. The data collection process is done through an open-source website 'Kaggle' in the form of a JSON file. When the collected data is completed, it is put for data preprocessing so that during the time of sentiment analysis, the data used is high-quality data. Thereafter, this will be followed by feature extraction where the word weights will be assigned to the different words which were used. Finally, the sentiment of customers' reviews will be assessed by implementing the Random Forest algorithm. This chapter contains the framework of the research as well as all the definitions in basic form and focuses on Exploratory Data Analysis (EDA).

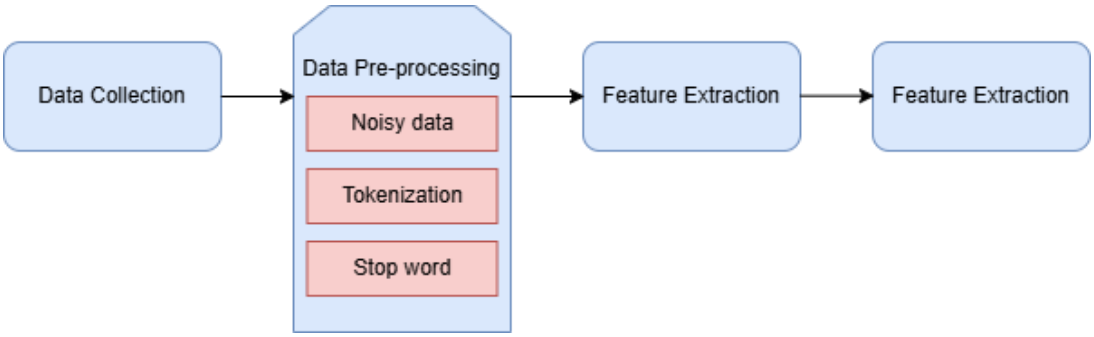


Figure 4.1 The conceptual framework of sentiment analysis on hotel review

Figure 4.1 depicts the conceptual structure for analyzing sentiments in hotel reviews. The workflow starts with gathering data, proceeds to data cleaning, then feature selection, and concludes with the implementation of a sentiment analysis technique. This framework acts as a roadmap for accurately classifying the sentiments expressed in the reviews. In this project, the sentiment analysis will distinguish reviews as either positive or negative.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is one of the methods to know in detail about the dataset. This includes the pattern, trends and relationship within the data. It is beneficial to understand of the data structure before implementing the machine learning algorithm. For this project, the data is collected from Kaggle. The original file is JSON. To ease the use of the dataset, the JSON file is converted to the CSV file. The JSON file is more than 9 GB, which makes it time-consuming to load. However, after conversion to a CSV file, the size is reduced to just 755,000 KB.

This difference occurs because JSON stores data in a hierarchical structure with significant overhead due to repetitive syntax, such as curly braces ({}), quotation marks, and keys for each data entry. While CSV only uses a simpler, flat structure without the additional syntax, resulting in a much smaller file size. To perform the Exploratory Data Analysis (EDA), Jupyter Notebook and Python were used throughout the process.

id	story
0	We went here with our kids for Xmas holiday and we really liked it. Large options of food for breakfast and lunch , you can really taste the quality of the food in there. The surrounding area is nice and clean. Good experience. Hardly recommended .
1	We have spent in this hotel our summer holidays both in summer 2014 and 2015- I was with my husband and my child (4 years old at present). I do really recommend this place- Staff si high qualified, Kind and really helpful- Animation staff get You involved, but always with discrection - Miniclub si super and activities offered are interesting and smart- Rooms clean, with AC and balcony- Restaurant offers a great selection of food - always. The beach si extremly closed to the hotel - Miniclub area offers some gazebos to have shade for kids- A lot of bicycles are available for free- I am completely satisfied of this hotel- Go in lime this!
2	I visited Hotel Baltic with my husband for some bike riding in the area, thinking it would just be another hotel. I was so wrong. We don't have children, but were so amazed at the attention to detail and kindness we experienced from every member of the staff. It was truly amazing.
3	I've travelled quite a numbers of hotels but this is the best place you can achieve with an excellent ratio quality/money. The equipe is really excellent. The restaurant's staff and the chef are perfect. Menu is always varying. Bar service is really fantastic. On the beach rather than in the hotel, anything is perfect and our holiday went like a dream. Although prices could seems quite high, you must consider that you could even forget your wollet at home. You'll never be required to spend any money. Kids are always happy and miniclub staff is really efficient. My daughter crying for our leaving could explain better what I'm writing.

Figure 4.2 The example of raw dataset from 1Lhotel-df.csv

Figure 4.2 illustrates an example of the raw dataset from the CSV file. The dataset consists of only two attributes which are id and story. This is because the primary focus is on the reviews, as the sentiment analysis relies on Natural Language Processing (NLP), which processes and analyzes the user reviews to determine their sentiment.

4.2.1 Data Cleaning

Data cleaning is a crucial stage in exploratory data analysis (EDA) that occurs prior to performing sentiment analysis. The purpose of data cleaning is to ensure the sentiment analysis is accurate. Figure 4.3 illustrates the number of missing values for the attributes id and story.

```
# check if dataset has null value
checknull = df.isnull().sum()
missing_data = pd.concat([checknull], keys=["Missing Values"], axis=1)
missing_data
```

	Missing Values
id	0
story	0

Figure 4.3 The number of missing values in the dataset

Figure 4.3 displays the number of missing values for each attribute. Both the "id" and "story" attributes have no missing or null values. It is important to check the missing values in the dataset to maintain consistency across all of the data and for better insight.

```
[45]: df['Char Count'] = df['story'].apply(len)
[47]: df["Word Count"] = df['story'].apply(lambda x: len(str(x).split()))
[48]: df['Sent Count'] = df['story'].apply(lambda x: len(nltk.sent_tokenize(str(x))))
[49]: df.head()
```

[49]:

	id	story	Char Count	Word Count	Sent Count
0	0	We went here with our kids for Xmas holiday an...	248	46	5
1	1	We have spent in this hotel our summer holiday...	645	116	3
2	2	I visited Hotel Baltic with my husband for som...	281	52	4
3	3	I've travelled quite a numbers of hotels but t...	641	109	10
4	4	We decided for this family holiday destination...	1335	259	4

Figure 4.4The number of characters, words and sentences for each review

Figure 4.4 displays the number of characters, words and sentences for each review before implementing the data pre-processing. Each review contains several sentences. The number of sentences is a crucial factor in sentiment analysis, as it reflects the complexity of

the sentiment expressed. For example, a review with a single sentence, such as “I love this hotel,” is straightforward, indicating a positive sentiment. However, a review consisting of multiple sentences, like “The hotel is good, I love the hospitality. But I hate the breakfast,” introduces more complexity, with both positive and negative sentiments. In this dataset, the minimum number of sentences is 1, and the maximum is 486.

	story	cleaned_story
0	We went here with our kids for Xmas holiday an...	we went here with our kids for xmas holiday an...
1	We have spent in this hotel our summer holiday...	we have spent in this hotel our summer holiday...
2	I visited Hotel Baltic with my husband for som...	i visited hotel baltic with my husband for som...
3	I've travelled quite a numbers of hotels but t...	ive travelled quite a numbers of hotels but th...
4	We decided for this family holiday destination...	we decided for this family holiday destination...
5	Great customer service and good restaurant ser...	great customer service and good restaurant ser...
6	This pousada is not too close to the downtown ...	this pousada is not too close to the downtown ...
7	Great hotel surrounded by nature! It was reall...	great hotel surrounded by nature it was really...
8	The property is surrounded by trees, which are...	the property is surrounded by trees which are ...
9	We really enjoyed our stay here, it was peacef...	we really enjoyed our stay here it was peacefu...

Figure 4.5 The clean data of data set

From Figure 4.5, the story column represents the raw data, while the cleaned_story column represents the processed data. The cleaned data has been converted to lowercase, with all numbers, punctuation, and extra whitespace removed. For example, in Figure 4.5, the word "I’ve" is transformed into "ive". Additionally, punctuation marks such as commas, exclamation marks, and full stops are eliminated during the cleaning process.

	cleaned_story	AfterStopWord
0	we went here with our kids for xmas holiday an...	went kids xmas holiday really liked large opti...
1	we have spent in this hotel our summer holiday...	spent hotel summer holidays summer husband chi...
2	i visited hotel baltic with my husband for som...	visited hotel baltic husband bike riding area ...
3	ive travelled quite a numbers of hotels but th...	ive travelled quite numbers hotels best place ...
4	we decided for this family holiday destination...	decided family holiday destination saw ranking...

Figure 4.6 The example of applying stop word

Figure 4.6 shows an example of a sentence before and after applying stopword removal. In the original sentence, there are many stopwords such as "I," "me," "myself," "we," "our," and others. The importance of removing stopwords lies in its ability to reduce noisy data. For instance, the original sentence ["I", "hated", "the", "food", "but", "room"] can be simplified to "hated food room" after removing stopwords. Words like "I," "the," and "but" do not contribute significant meaning for sentiment analysis. Instead, sentiment analysis focuses on the more meaningful words, such as "hated," "food," and "room," which carry the actual sentiment of the review. By eliminating stopwords, we reduce unnecessary complexity in the data, allowing the sentiment analysis model to focus on the key elements that reflect user opinions

	AfterStopWord	split_review_clean
0	went kids xmas holiday really liked large opti...	[went, kids, xmas, holiday, really, liked, lar...
1	spent hotel summer holidays summer husband chi...	[spent, hotel, summer, holidays, summer, husba...
2	visited hotel baltic husband bike riding area ...	[visited, hotel, baltic, husband, bike, riding...
3	ive travelled quite numbers hotels best place ...	[ive, travelled, quite, numbers, hotels, best,...
4	decided family holiday destination saw ranking...	[decided, family, holiday, destination, saw, r...

Figure 4.7 The tokenization of review

Figure 4.7 illustrates the process of splitting words, commonly known as tokenization. This process utilizes the split() method in Python. Tokenization is a crucial step as it transforms a full sentence into individual units, or tokens, which can then be analyzed independently. Tokenization helps extract meaningful features from text. For instance, in the example [great, customer, service, and, good, restaurant], the word “good” conveys a strong positive sentiment. Identifying such words is essential for sentiment analysis. Furthermore, tokenization prepares the data for subsequent processing in machine learning models, enabling more accurate predictions and insights.

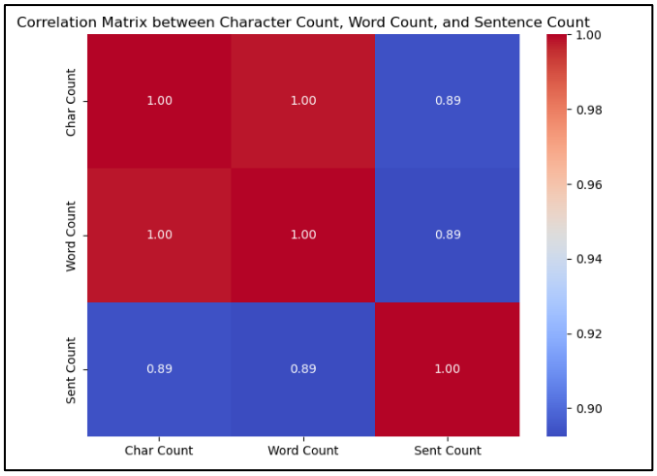


Figure 4.8 The correlation between character, word and sentence count

Figure 4.8 displays the correlation between character, word and sentence count. The nearest value to 1 indicates the stronger the positive correlation. Based on the figure 4.8, the character count and word count strongly correlate positively. While character count and sentence count have moderate correlation that depends on sentence length. Similar to word count and sentence count that have moderate correlation that influenced by sentence structure.

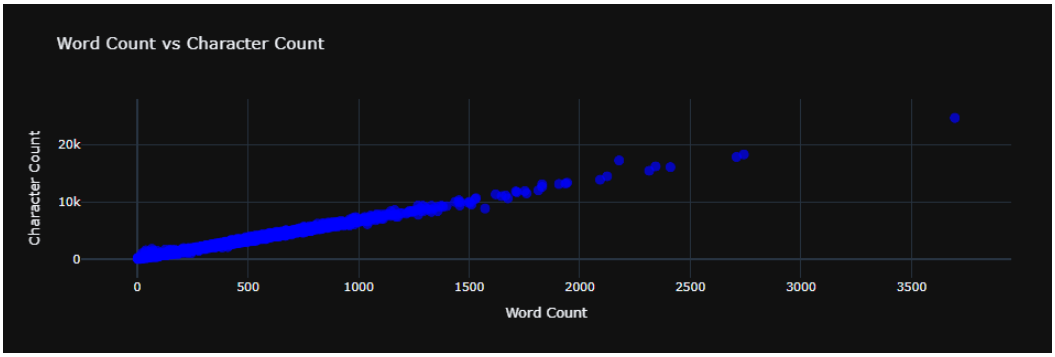


Figure 4.9 The scatter plot of word count and character count

Figure 4.9 illustrates an upward trend, showing a positive correlation between word count and character count. Reviews with more words tend to have a higher character count, which is expected. The scatter plot displays a linear pattern, indicating that each additional words in a review consistently adds a similar number of characters. Additionally, the graph

shows a few points on the far-right side with exceptionally high word and character counts, suggesting potential outliers that may require further inspection. Most of the data points are clustered in the lower-left region of the graph, signifying that the majority of reviews are short or concise, while longer reviews are relatively uncommon in the dataset.

4.2.2 Feature Engineering

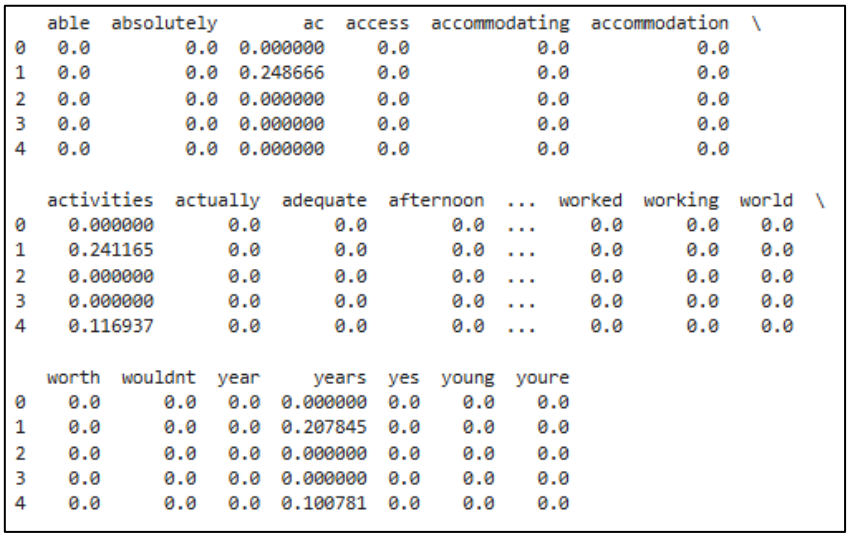
In this project, TF-IDF and K-Mean Clustering is used to the feature extracting for the sentiment analysis. The TF-IDF is a process that turns all the words into the numerical value. After calculating the TF-IDF score, it will be clustered using the K-Mean Clustering. The number of clusters is 2 which represents positive and negative sentiment.

TF-IDF is used to evaluate how important a word is to a document within a collection or corpus. In this project, the words is vectorized to the top 500 most important words which means, the TF-IDF will calculate the top 500 words that always being used by the user. Then, the score of each word is based on these 500 words. Figure 4.10 illustrates few of 500 words that commonly used.

```
['able' 'absolutely' 'ac' 'access' 'accommodating' 'accommodation'
 'activities' 'actually' 'adequate' 'afternoon' 'air' 'airport' 'amazing'
 'amenities' 'apartment' 'area' 'areas' 'arrival' 'arrived' 'ask' 'asked'
 'ate' 'atmosphere' 'attentive' 'available' 'average' 'away' 'awesome'
 'bad' 'balcony' 'bar' 'bars' 'basic' 'bath' 'bathroom' 'bathrooms'
 'beach' 'beautiful' 'bed' 'bedroom' 'beds' 'best' 'better' 'big' 'bit'
 'book' 'booked' 'booking' 'break' 'breakfast' 'breakfasts' 'brilliant'
 'bring' 'buffet' 'building' 'bus' 'business' 'busy' 'called' 'came' 'car'
 'card' 'care' 'center' 'central' 'centre' 'certainly' 'chairs' 'change'
 'charge' 'cheap' 'check' 'checked' 'checkin' 'children' 'choice'
 'choices' 'choose' 'chose' 'city' 'clean' 'cleaned' 'cleaning' 'close'
 'club' 'coffee' 'cold' 'come' 'comfortable' 'comfy' 'coming'
 'complimentary' 'continental' 'convenient' 'cooked' 'cool' 'cost'
 'couple' 'course' 'customer' 'daily' 'day' 'days' 'deal' 'decent'
 'decided' 'decor' 'decorated' 'definitely' 'delicious' 'desk' 'didnt'
 'different' 'dining' 'dinner' 'dirty' 'disappointed' 'distance' 'dont'
 'door' 'double' 'downtown' 'drink' 'drinks' 'drive' 'early' 'easily'
 'easy' 'eat' 'efficient' 'eggs' 'end' 'english' 'enjoy' 'enjoyable'
 'enjoyed' 'entertainment' 'entire' 'especially' 'evening' 'excellent'
 'expect' 'expected' 'expensive' 'experience' 'extra' 'extremely'
 'fabulous' 'facilities' 'fact' 'family' 'fantastic' 'far' 'fault' 'feel'
 'felt' 'fine' 'flight' 'floor' 'food' 'free' 'fresh' 'fridge' 'friendly'
 'friends' 'fruit' 'fun' 'garden' 'gave' 'getting' 'given' 'glass' 'going'
 'good' 'got' 'grand' 'great' 'greeted' 'grounds' 'group' 'guest' 'guests'
 'gym' 'half' 'happy' 'hard' 'hear' 'help' 'helped' 'helpful' 'high'
 'highly' 'hilton' 'holiday' 'home' 'hope' 'hot' 'hotel' 'hotels' 'hour'
```

Figure 4.10 The words of data feature

Figure 4.10 shows the words that will be used to calculate the score of the reviews. The words are able, absolutely, access, accommodating, activities and others. These scores indicate how important each word is for each document within the given corpus.



	able	absolutely	ac	access	accommodating	accommodation	\
0	0.0	0.0	0.000000	0.0	0.0	0.0	
1	0.0	0.0	0.248666	0.0	0.0	0.0	
2	0.0	0.0	0.000000	0.0	0.0	0.0	
3	0.0	0.0	0.000000	0.0	0.0	0.0	
4	0.0	0.0	0.000000	0.0	0.0	0.0	

	activities	actually	adequate	afternoon	...	worked	working	world	\
0	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	
1	0.241165	0.0	0.0	0.0	...	0.0	0.0	0.0	
2	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	
3	0.000000	0.0	0.0	0.0	...	0.0	0.0	0.0	
4	0.116937	0.0	0.0	0.0	...	0.0	0.0	0.0	

	worth	wouldnt	year	years	yes	young	youre
0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
1	0.0	0.0	0.0	0.207845	0.0	0.0	0.0
2	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
3	0.0	0.0	0.0	0.000000	0.0	0.0	0.0
4	0.0	0.0	0.0	0.100781	0.0	0.0	0.0

4.11 The score for the first 5 rows

Figure 4.11 shows the score of the first 5 rows. It is for the second row, the score for “ac” is 0.24866 because the “ac” is appeared in the reviews. Same goes to activities, the score is 2.41165. The review after clean is “spent hotel summer holidays summer husband child years old present really recommend place staff si high qualified kind really helpful animation staff get involved always discrection miniclub si super activities offered interesting smart rooms clean ac balcony restaurant offers great selection food always beach si extremly closed hotel miniclub area offers gazebos shade kids lot bicycles available free completely satisfied hotel go lime”.

K-Means clustering is a widely used algorithm for grouping data points into clusters based on their similarity. In this project the number of clusters is 2 which represent positive and negative sentiment. K-Means will calculate the average position of all the data points assigned of the clusters based on the TF-IDF vectors.

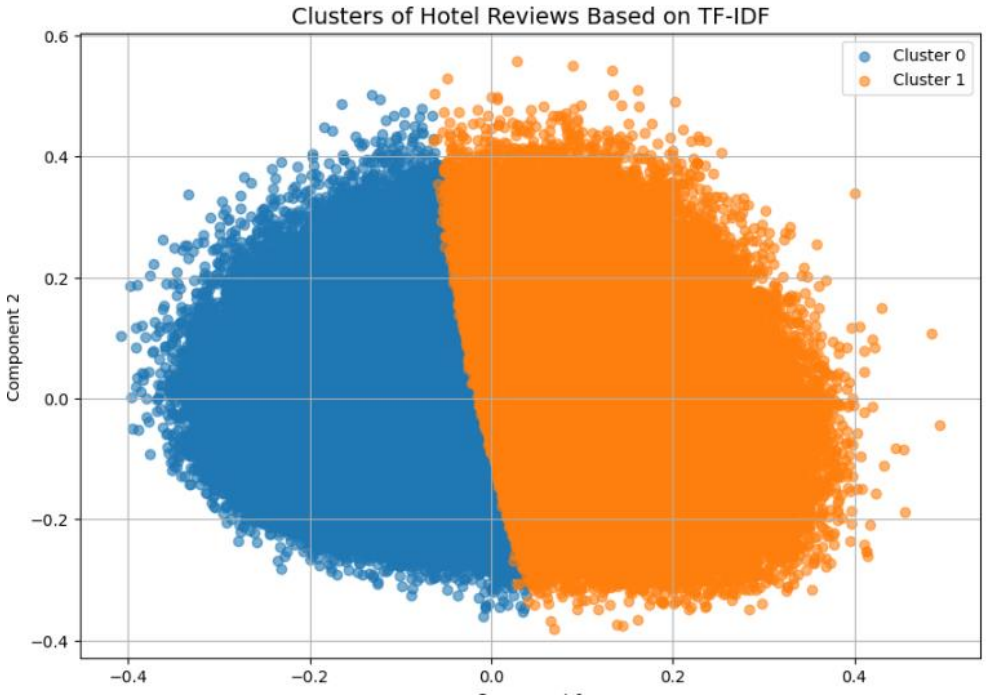


Figure 4.12 The proportion of clustering

Figure 4.12 illustrates the proportion of two distinct groups, represented by the blue and orange colors. The blue represents the cluster 0 while the orange represents the cluster 1. In this cluster, 0 represent positive and 1 represent negative. There is significant overlap between the clusters that indicate the reviews in both clusters are using similar in terms of the words used. Apart from that, it can also be seen an outlier. It may happen because it does not fit well into either of the clusters. The content of the review may contain unique terms not common across other reviews.