SENTIMENT ANALYSIS ON HOTEL REVIEWS USING MACHINE LEARNING

NURFATINI ATIQAH BINTI HAMIDI

UNIVERSITI TEKNOLOGI MALAYSIA

CHAPTER 1

INTRODUCTION

1.1 Introduction

As a traveller, booking a hotel is part of the planning. Usually, other people's review will influence a user's decision to find and book the best hotel to stay. Reviews available on the internet are more relevant, actual and detailed than the reviews found in hotel brochures (Walter Kasper & Mihaela Vela, 2012). There are various sources of online platform to find the reviews to get a better insight about the hotel's reputation. For example, Google Review, Agoda and Booking.com. So, customer's reviews play an important part for both user and business owner for decision making and improvement for the services.

Sentiment analysis also known as "opinion mining" or "emotion AI" is a technique used to gather and examine user thoughts, opinions, and responses to a particular topic (Zahid & Linköping University, 2020). Sentiment analysis is frequently performed using text mining with Natural Language Processing (NLP) tools to examine evaluations and reactions (Zahid & Linköping University, 2020). It involves analysing the review in form of text to determine whether it expresses positive, negative, or neutral sentiment.

There is a variety of machine learning algorithms such as Naive Bayes, Logistic Regression, Support Vector Machines (SVM) and Random Forests that can be employed to the categorization of the sentiment analysis of hotel reviews. In this project, TF-IDF calculation is used. TF stands for term frequency and IDF stands for inverse document frequency. In a simple word, it provides those keywords, using which some specific documents can be identified or categorized. Random Forest is a popular machine learning algorithm that can be used for both classification and regression problems in Machine Learning.

1.2 Problem Background

In this digital era, online reviews have become a key factor influencing consumer decisions making in the hospitality industry. User rely much based on reviews from previous customers to gauge the hotel quality. The review also becomes a vital aspect of a hotel's online reputation and competitive edge. Platforms such as TripAdvisor, Agoda, and Booking.com received thousands of hotel reviews from their customer that reflecting the customer's experience and emotion towards their stay. This huge amount of data contains crucial opinion related information that can be used to benefit for businesses and other aspects of commercial and scientific industries (Rezwanul et al., 2017). To manually tracking and pulling out all the data is challenging. Traditional methods are very time-consuming and unable to capture the depth and exact details of customer sentiments expressed in text form. Customer often convey mixed emotions within a single review, such as positive remarks about room quality paired with complaints about service. Thus, sentiment analysis is one of the techniques that can be used to capture and interpret the review into simple categorization.

1.3 Problem Statement

Hotel reviews have become an important thing to the business owner because it can be source of income to the business. However, due to the high volume of reviews generated on daily basis, it is difficult to the business owner to respond in a timely manner. Traditional analysis method is very time consuming. For example, business owners need to read one by one the mix of good and bad reviews. It can lead to misleading the details from the cust

Apart from that, the inconsistency between review and rating from the customer, customers give a good review but rate for one star. For example, the review "Good place to stay" but the rating is one star that represents not good. By implement the sentiment analysis, it can help the business owner to give accurate interprets of the emotion and satisfaction levels from the customer reviews.

1.4 Research Objectives

The proposed project aims to achieve the following objectives:

(a) To conduct exploratory data analysis to identify patterns of hotel reviews.

(b) To design and implement sentiment analysis that predict the review either positive or negative.

(c) To conduct comprehensive evaluations on the develops predictive model and build an interactive dashboard.

1.5 Scope of Study

The proposed of this project is sentiment analysis on Hotel Review. The sentiment analysis prediction based on the customer review (text) and the prediction is being categorize into two types which are positive and negative. In this project, the dataset used downloaded from Kaggle and word cloud will be used to interpret the review. The dashboard will be developed using data visualization tools in PowerBI to provide an insight of the positive and negative review.

CHAPTER 2

LITERATURE REVIEW

2.1 Sentiment Analysis

Sentiment refers to a thought, belief, or notion that comes from one's feelings regarding a situation or a perspective on a particular issue. On the other hand, analysis involves the process of closely studying or examining something to gain a deeper understanding or form an opinion and judgement based on that examination. Sentiment Analysis is a Natural Language Processing (NLP) task designed to identify sentiments and opinions within written content.(Birjali et al., 2021) Opinion and sentiment can be classified into three main types which are regular opinions, comparative opinion and suggestive opinions. Regular opinions are about one thing, suggestive opinions are about one or more things, and comparative opinions are about comparing or contrasting more than one thing. (Shayaa et al., 2018)

In the age of modern science, everything is based on online and on the internet. Social media's rapid expansion has made public opinions and thoughts more accessible, which makes sentiment analysis an essential tool for understanding public opinion in a variety of fields, including politics and commerce. (Tan et al., 2023) For example, when tourists want to choose a comfortable hotel for their trip, they will look for reviews from other travellers. The reviews in the internet are more trusted than in the brochure because the user’s review can be verified. Because such sentiments and opinions play a crucial role in our daily lives, making it essential to examine this user-generated data to effectively monitor public opinion and aid in decision-making. (Birjali et al., 2021)

Consequently, the area of sentiment analysis has attracted greater attention in the past fifteen years among research groups. Since 2004, sentiment analysis has become the most rapidly growing and vibrant area of research, as evidenced by a notable increase in the number

of papers focused on sentiment analysis recently. (Mäntylä et al., 2018) Figure 2.1 illustrates the increasing interest in sentiment analysis as indicated by Google Trends.

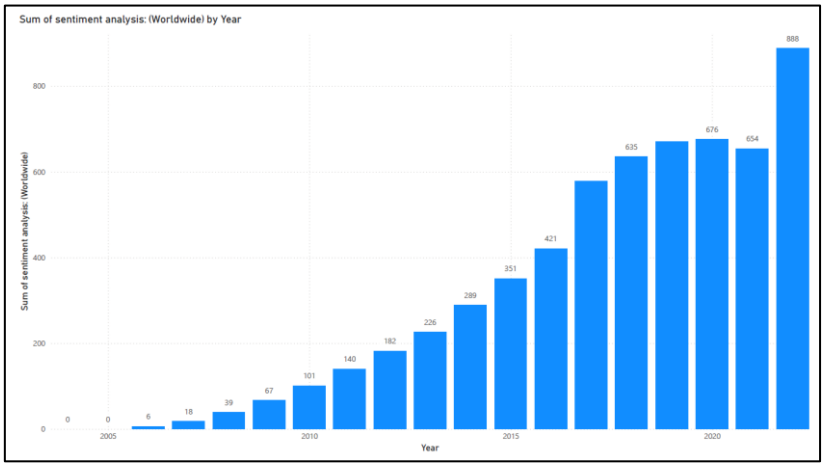


Figure 2.1 Trends of sentiment analysis according to Google Trends(trends.google.com/trends)

2.2 Sentiment Analysis Levels

Sentiment Analysis has been explored at various levels, including Document Level, Sentence Level, Phrase Level, and Aspect Level. Figure 2.2 illustrates sentiment analysis at various levels, including document, sentence, phrase, and aspect. Sentiment analysis at the document level evaluates an entire document to provide a single polarity classification. This approach assigns an overall sentiment label, such as positive, negative or neutral to the complete document. Supervised and unsupervised learning methods can be use at every of the stage to categorize the any words or text. (Bhatia et al., 2015)

Sentence level of sentiment analysis is analyse each sentence and find with corresponding polarity. It focuses on determining the sentiment by individual sentences to positive, negative or neutral. This is very useful when a document covers a wide range of associated emotions. (Yang & Cardie, 2014) The polarity of each sentence will be assessed separately using the same techniques as at the document level, but with enhanced training data and upgraded processing capabilities. (Wankhade et al., 2022) For example, consider a hotel review “The battery life is excellent. However, the food is not good”. So, it will analyse

sentence by sentence. “The battery life is excellent” is positive and “However, the food is not good” is negative.

Sentiment analysis at the phrase level can additionally be conducted by identifying opinion words at that specific level, which is then followed by a classification process. (Wankhade et al., 2022) This could be beneficial when reviewing products across several lines; it is noted that one element is conveyed in a single phrase. (Thet et al. 2010). It goes deeper than sentence-level sentiment analysis by examining smaller linguistic units, such as clauses or phrases, and labelling their sentiment as positive, negative or neutral. For example, “The food was excellent but the service was terrible”. It will analyse by phrase which “The food was excellent” is positive “but the service was terrible" is negative.

Lastly, aspect level is where sentiment analysis is performed. It focuses on identifying sentiment associated with specific aspects or attributes of an entity, rather than analyzing the sentiment of an entire document, sentence, or phrase. The emphasis is on each component within the sentence, with polarity assigned to every element, followed by the calculation of an overall sentiment for the complete sentence. (Schouten & Frasincar, 2015) For example, the sentence is “The display is beautiful, but the sound quality is disappointing, and the battery life is average.”, the aspect extraction is display, sound quality and battery life. The sentiment classification for display is positive, sound quality is negative and battery life is neutral.

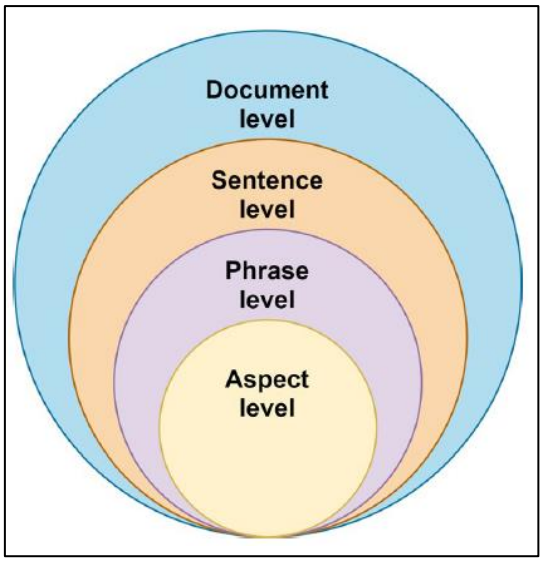


Figure 2.2 Sentiment analysis level (Wankhade et al., 2022)

2.3 Sentiment Analysis Approach

There is a multiple of sentiment analysis approach. For example, Lexicon-based approach, machine learning approach and hybrid approach. Figure 2.3 shows the example of sentiment analysis approach.

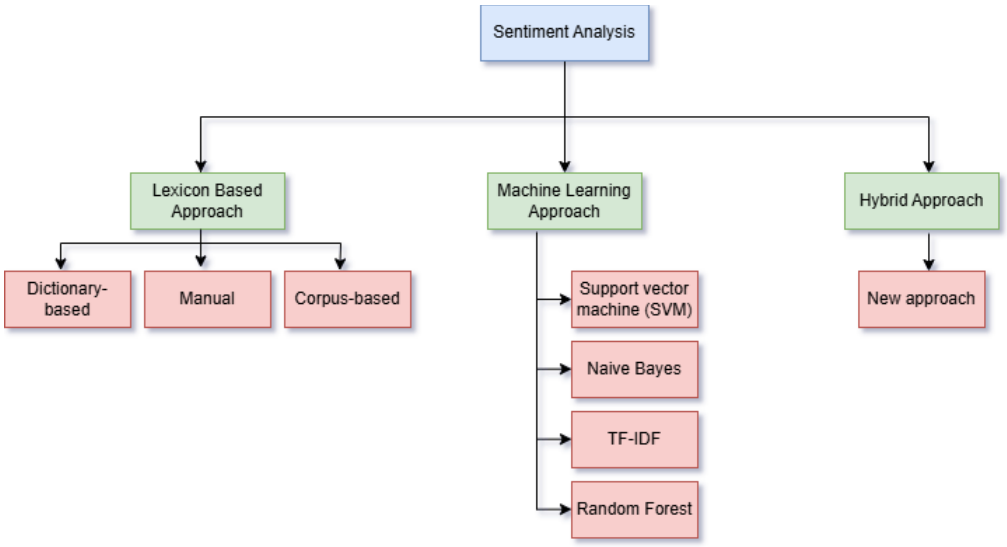


Figure 2.3 Sentiment analysis approach

Figure 2.3 shows the sentiment analysis approach. Sentiment analysis has a few approaches which are Machine Learning, Lexicon-based and Hybrid. Machine learning based approach uses classification technique to classify text. It comprises two groups of documents which are a training set and a test set. The training set is utilized to grasp the distinguishing features of a document, whereas the test set is employed to evaluate the classifier's performance. (Alessia et al., 2015) Support Vector Machine (SVM) is a popular technique used to recognize classification algorithm and it is based on supervised machine learning. (Khairnar & Kinikar, 2013). Naïve Bayes (NB) is an elementary classifier and is among the most frequently utilized algorithms in text classification. (Birjali et al., 2021) Random Forest focuses on improving and retaining classification trees. (Breiman, 2001), and Term Frequency – Inverse Document Frequency (TF-IDF).

The lexicon-based method involves gathering tokens, with each token linked to a specific score that reflects its neutral, positive, or negative characteristics in the text.

(Kiritchenko et al., 2014) Based on (Wankhade et al., 2022), there are three methods for developing a sentiment lexicon such as manual creation, corpus-based techniques, and dictionary-based approaches. The manual creation method is quite challenging and demands a considerable of time. It will take a lot of time to manually applied this technique. Corpus-based techniques are capable of generating opinion or reviews in the form of words with result with high precision. In dictionary-based methods, it starts with compiling a small collection of manually curated opinion words from various platforms that have established orientations. Then, it will be expanded by looking for contrast and equivalent words in the WordNet or WordCloud dictionary.

Lastly, Hybrid approach is the addition of machine learning and Lexicon-based approach. The primary purpose of the hybrid approach is to gain high accuracy from machine learning while maintaining stability from the lexicon-based method. (Birjali et al., 2021) The focus of this research is more into machine learning. For example, (Amrani et al., 2018) suggested a hybrid method utilizing machine learning that incorporates both Random Forest (RF) and Support Vector Machine (SVM). Their findings highlighted that the standalone models of SVM and RF achieved accuracies between 81.01% and 82.03%. While the combined hybrid model reached an accuracy of nearly 84% based on the product review from Amazon.com. Table 2.1 shows the possible advantages and limitation of sentiment analysis approach.

Type of approach	Advantages	Limitation
Machine Learning	The capability to adjust and develop trained models for particular uses and situations.	The limited applicability to new data arises from the need for labelled data, which can be expensive or even unaffordable.
Lexicon-based	Broader term inclusion, annotated data, and the process of learning are not necessary.	A restricted set of words in the lexicons is assigned a particular sentiment orientation and score for each word
Hybrid	Lexicon/learning symbiosis involves identifying and gauging sentiment at the	Noisy reviews

	conceptual level, along with reduced sensitivity to shifts in topic domain.	

Table 2.1 The Advantages and Limitation of Sentiment Analysis Approach

2.4 Machine Learning Approach

Sentiment analysis is more effectively performed using machine learning, although developing precise machine learning models can be challenging. (Khomsah, 2020) Various strategies are available to improve machine learning models, including techniques, evolutionary algorithms in this field, and intelligent swarms. (Rizaldy & Santoso, 2017) The machine learning algorithm TF-IDF, Random Forest, TF-IDF and Naïve Bayes will be discussed.

2.4.1 The Term Frequency-Inverse Document Frequency (TF-IDF)

The term frequency-inverse document frequency, commonly referred to as TF-IDF, is a prevalent technique used to assess the significance of a word within a document. TF measures how often a specific term appears within the document. Since documents can differ significantly in length, a term will probably show up much more often in longer documents compared to shorter ones. (Wankhade et al., 2022)

$$Term\ Frequency = \frac{(Number\ of\ times\ term\ t\ present\ in\ a\ document)}{(Total\ number\ of\ terms\ in\ the\ document)}$$

IDF (Inverse Document Frequency) is used to give lower weight to words that occur frequently and to give larger words to words that occur rarely. (Qaiser & Ali, 2018) There are certain words such as “is,” “an,” “and,” “when,” and others that appear often but carry little significance. IDF is calculated as IDF (t) = log(N/DF), where N is the number of documents and DF is the number of document containing term t.(Ahuja et al., 2019)

2.4.2 Random Forest

Random forest is an ensemble technique that generates multiple decision trees, which are then combined into a forest. (Zahid-samza595, 2020) Each tree in the Random Forest makes a class prediction and the Random Forest determines the outcome based on the majority of votes. (Saad & Aref, 2020) A random forest can be described as a collection of tree-structured classifiers {h(x, Ok), k=1,…}, where {Ok} represents independently and identically distributed random vectors. Each tree in this ensemble gives a single vote for the most common and prominent class associated with the input x.(Ahmad et al., 2017) Every tree is constructed using a bootstrap sample of the data and the predictions from all trees are ultimately combined through majority voting. (Shayaa et al., 2018)

2.4.3 Support Vector Machine (SVM)

The Support Vector Machine or SVM is a widely recognized and highly effective method for classification in machine learning. (Saad & Aref, 2020) Support Vector Machines (SVM) represent as non-probability of statistic and supervised learning techniques frequently used for classification tasks. The primary objective of SVM is to determine the hyperplane that best separates the data into distinct categories. (Wankhade et al., 2022) Text classification with SVM involves representing each document as a vector, where the dimensions correspond to the count of unique keywords utilized in the training data. (Zahid-samza595, 2020) SVM can be used for data that is linearly separable, and it can also handle non-linear data with the appropriate kernel functions or techniques. (Bania, 2020) Figure 2.4 The diagram demonstrates the concept of SVM. On the left are the original items, while the right displays them after being rearranged through a mathematical function called a kernel, which is referred to as mapping or transformation. Once the transformation occurs, the rearranged items can be separated linearly, thus eliminating the need for complex curves to distinguish between the objects. (Devika et al., 2016)

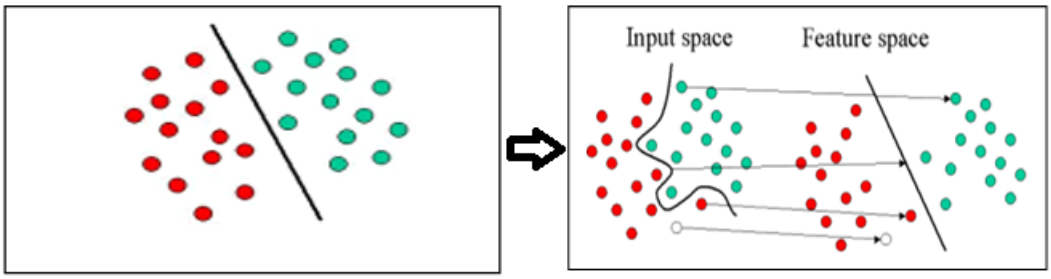


Figure 2.4 SVM illustration (Devika et al., 2016)

2.4.4 Naïve Bayes (NB)

Naïve Bayes (NB) is an effective classification algorithm used for categorizing data based on probability analysis. (Ahuja et al., 2019) In machine learning it is in family of sample probabilistic classifier based on Bayes theorem and depends on BoW feature extraction. (Devika et al., 2016) It typically involves a small data set for training and is subsequently utilized to estimate the parameters required for classification objectives. (Shayaa et al., 2018) It operates effectively with extensive datasets, being quick and precise while maintaining a minimal computational expense. (Saad & Aref, 2020) Figure 2.5 shows the Bayes Theorem used in NB.

$$P(c|D) = \frac{P(D|c) * P(c)}{P(D)}$$

Figure 2.5 Bayes Theorem (Zahid-samza595, 2020)

2.5 Review of Similar Works

Table 2.2 shows similar works that had been done by previous researchers about the project. All of the researchers use the same features which is the term frequency-inverse document frequency (TF-IDF). The different is the classifier. Makhmudah et al and Alzyout et al use the same classifier which is Support Vector Machine. The accuracy for dataset used by

Makhmudah et al is the highest which is 99.5% while Alzyout et al used self-collected dataset has the accuracy 78.25%. Alsalman that used Multinomial Naïve Bayes has the second highest accuracy which is 87.5%. While Rathi et al used AdaBoost which unsupervised method, AdaBoost has the lowest accuracy which is 67%.

Author	Dataset	Features	Classifier	Accuracy (%)
(Rathi et al., 2018)	Sentiment140, Polarity Dataset, and University of Michigan dataset	TF-IDF	AdaBoost	67
(Makhmudah et al., 2019)	Tweets related to homosexuals	TF-IDF	Support Vector Machine (SVM)	99.5
(Gupta et al., 2019)	Sentiment140	TF-IDF	Neural Network	80
(Alsalman, 2020)	Arabic Tweets	TF-IDF	Multinomial Naïve Bayes	87.5
(Alzyout et al., 2021)	Self-collected dataset	TF-IDF	Support Vector Machine (SVM)	78.25

Table 2.2 Example of similar works

2.6 Research Gap

Firstly, the research not enough attention to the other method than Machine Learning. For example, Lexicon-based approach. There are few Lexicon-based approach techniques that widely used to do the classification. This is because the course is more focus on the Machine Learning. Apart from that, the research also not enough attention on the project that the dataset is in other languages. For example, in this project sentiment analysis on hotel reviews. So, the reviews may be in Arabic or Germany language. But, for this research only focus more on the English and eliminate anything other than English language.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The project methodology involves four steps which are data collection, data preprocessing, feature extraction, sentiment algorithm, and evaluation. For the data collection, the data will be downloaded from an open source. For data preprocessing, it is carried out to eliminate noise and address inconsistent data.. For feature extraction, it is carried out as word weighing. Lastly, for sentiment algorithm, it is carried out to classify the review as positive or negative.

3.1.1 Proposed Method

There are five steps that need to be followed to conduct comprehensive evaluations on the developed predictive model and build an interactive dashboard of sentiment analysis on hotel review using machine learning which are data collection, data preprocessing, feature extraction, sentiment algorithm and evaluation. Figure 3.1 shows all of the four steps.

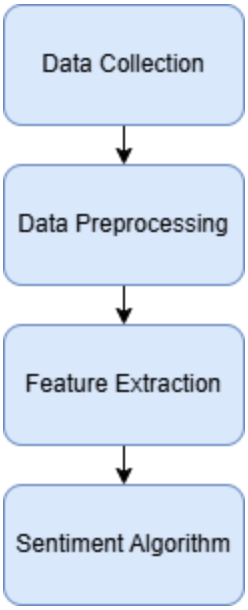


Figure 3.1 The steps of sentiment analysis

Phase	Objective	Activities
Data collection	• To identify the attributes of sentiment analysis on hotel review	• Download dataset from an open source
Data Preprocessing	• To eliminate noise and address inconsistent data.	• Normalization • Tokenization • Lemmatization
Feature extraction	• To evaluate the word weighing	• TF-IDF
Sentiment algorithm	• To classify the sentences as positive or negative	• Random Forest • Support Vector Machine
Evaluation	• To evaluate the accuracy of the sentiment algorithm	•

Table 3.1 The details of project development

3.2 Data Collection

Data collection of sentiment analysis of hotel reviews using machine learning is extracted from the open source which is Kaggle. The size of the dataset is 9,835,858 KB. It has

two attributes which are ID and Story. ID stands for the continuous number starting from one. Story stands from the sentence of reviews from the users. Figure 3.2 shows the number of rows and column in the dataset.

Figure 4.3 The number of rows and columns in the dataset

Figure 4.3 shows the number of rows and columns in the dataset. The dataset contains 1,100,001 rows and 2 attributes (columns). The data type for the "id" attribute is int64, which stores numeric data, while the "story" attribute has the data type object, which is used to store string data or mixed types.

3.3 Data Pre-processing

Data Pre-processing is a process of cleaning the raw data to become structured data. Online reviews are usually not clean and contain a lot of special characters, emoticons, URLs, hashtags and other text that are not necessary. Data pre-processing is one of the steps to ensure the high-quality data is used for the process of sentiment analysis. The process includes normalization, tokenization and lemmatization.

3.3.1 Normalization

Normalization is an important aspect in text analysis. It ensures that the data is consistent and comparable. Normalization can clean data in multiple ways. The procedure consists of converting all text to either all uppercase or all lowercase, eliminating punctuation, and transforming numbers into their written word forms. Apart from that, it can also remove redundant or duplicate data and handle missing values. For example, if there is a column with missing data, it can replace the missing data with 0 or the mean of the column.

Before Normalization	After Normalization
The Hotel is great. I give 4 stars and will come back again!!!!	the hotel is great i give four stars and will come back again
This hotel is AWESOME ❤	this hotel is awesome

Figure 3.2 Example after normalization

Figure 3.2 shows before and after normalization. The second sentence, “This hotel is awesome ❤” is raw data, and when it is normalized, the emoticon is removed, resulting in all the sentences becoming lowercase, “this hotel is awesome.”

3.3.2 Tokenization

Tokenization is the process of cutting the input string based on each compiler word. (Farisi et al., 2019) In a simple word, it separates the sentences into each word which is referred to as a token. It is a fundamental step in natural language processing (NLP).

Normalization	Tokenization
the hotel is great i give four stars and will come back again	['the', 'hotel', 'is', 'great', 'i', 'give', 'four', 'stars', 'and', 'will', 'come', 'back', 'again']
this hotel is awesome	['this', 'hotel', 'is', 'awesome']

Figure 3.3 Example of tokenization

Figure 3.3 shows how tokenization works. It is breaking down the sentences into words as a token. The second sentences “this hotel is awesome” is breaking down into four words which are ['this', 'hotel', 'is', 'awesome'].

3.3.3 Lemmatization

Lemmatization involves transforming a word into its base form for every word that has been tokenized. By employing lemmatization, every prefix and suffix is stripped from each word, converting them into their base forms to improve efficiency in text processing. For example, "running" will be transformed into "run". Examples of another word that will be taken out during lemmatization are “were”, “and”, “an”, “are” and others.

Original	Lemmatization
The geese are flying towards the mountains and running fast.	the goose fly towards the mountain run fast

Figure 3.4 Example of lemmatization

Figure 3.4 shows the lemmatization result. The raw sentence is “The geese are flying towards the mountains and running fast.” After lemmatization, the sentence becomes “the goose fly towards the mountain and run fast,” where flying and running are present participles that change to the base form fly and run.

3.4 Feature Extraction

Feature extraction is one of the essential steps after data pre-processing. Feature extraction has multiple techniques that can be used. In this project, TF-IDF will be used to calculate the weight of the sentence. TF-IDF is one of the most used techniques for text extraction. TF-IDF transforms the review text data into numbers. The steps in calculating the TF-IDF as follows:

1) 1) Determining the frequency of each word's Term Frequency (TF).

- The number of Term Frequency (TF) is calculated by separating sentences into one word and each word is given a value of 1.

2) Calculating the frequency of documents (DF) for each word.

- The document frequency (DF) is calculated by adding up the TF values for each word.

3) Determine the value of inverse document frequency (IDF).

$$IDF(w) = \log(\frac{N}{DF}) \quad (3.1)$$

Where N is the number of documents and DF is the number of documents containing the term t.

4) Determine the weight by multiplying the TF value with the IDF.

$$W_{ij= tf_{ij} \log(\frac{D}{df_j})} \quad (3.2)$$

For instance, consider a document that consists of 100 words, with the word “happy” occurring 10 times. In this case, the term frequency would be calculated as 10/100=0.1. Now, let's assume there are 50000 documents in total, and only 500 of those contain the word “happy.” Therefore, the IDF (happy) can be expressed as 50000/500=100, resulting in log(100) = 2. Consequently, the TF-IDF (happy) would be 0.1*2= 0.2

3.5 Sentiment Algorithm

The sentiment algorithm used for this project is the K-Means Cluster. It is an ensemble of decision tree algorithms that can be used for both classification and regression. In this algorithm generally, more trees correspond to better performance and efficiency. In a given training set, extract a sample set of data points by using the bootstrap method. After this construct a decision tree based on the output of the previous step. Apply the previous two steps and we will get the number of trees.

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).

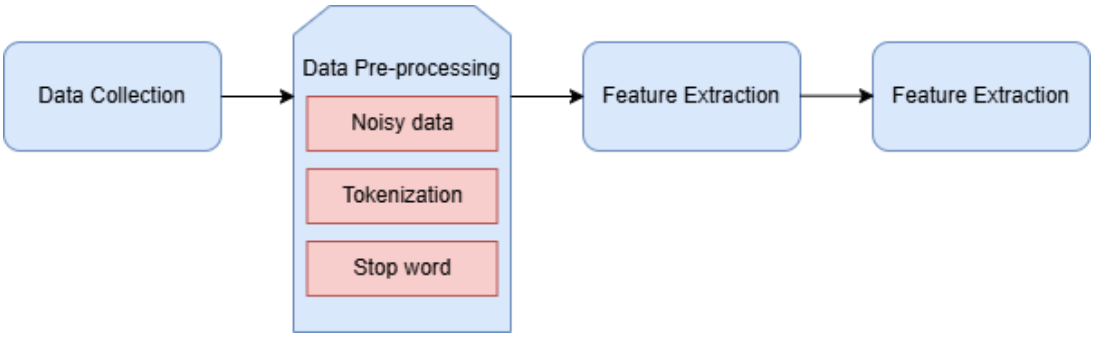


Figure 4.1 The conceptual framework of sentiment analysis on hotel review

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.

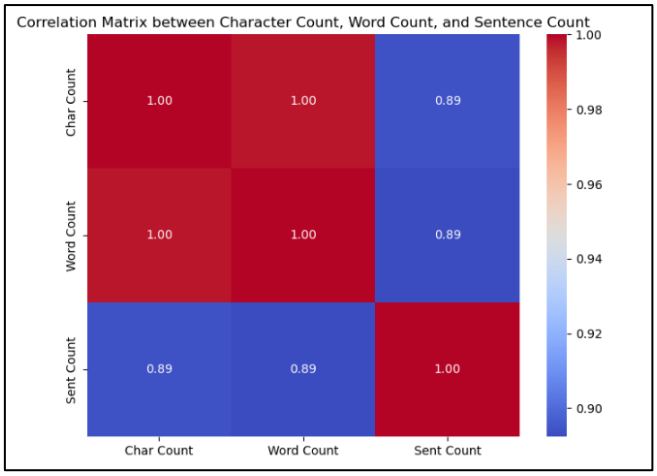


Figure 4.8 The correlation between character, word and sentence count

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.

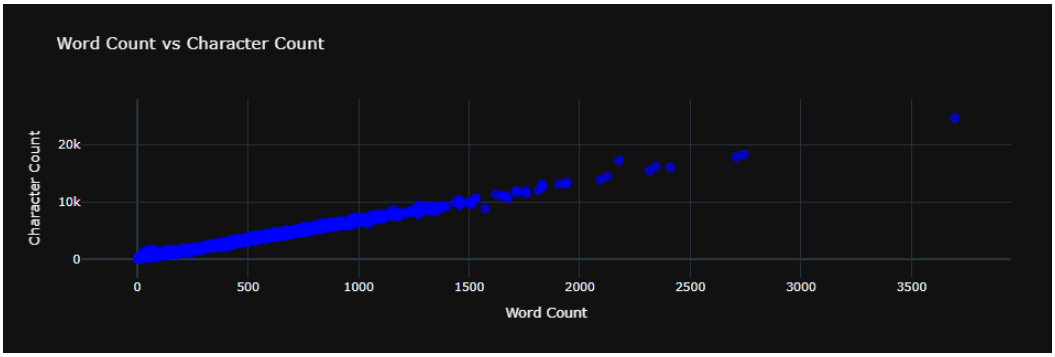


Figure 4.9 The scatter plot of word count and character count

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

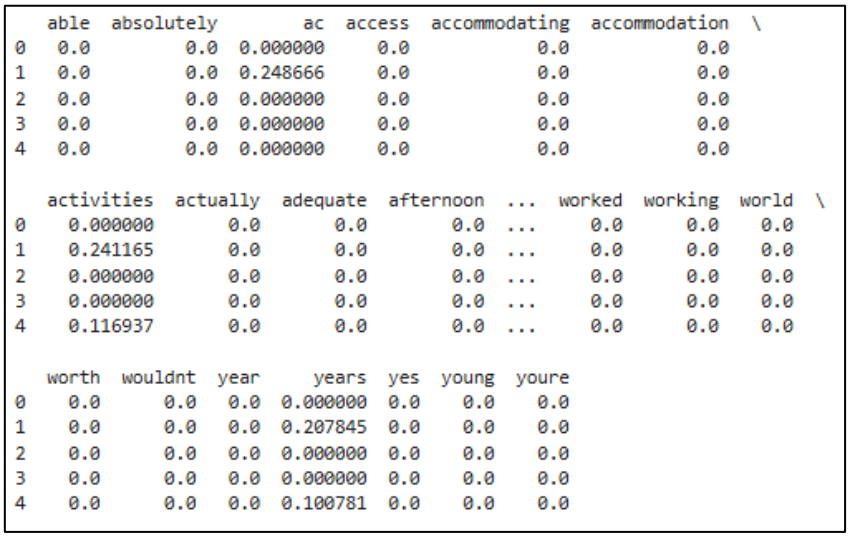
In this project, TF-IDF and K-Mean Clustering is used to the feature extracting for the sentiment analysis. The TF-IDF is a process that turns all the words into the numerical value. After calculating the TF-IDF score, it will be clustered using the K-Mean Clustering. The number of clusters is 2 which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

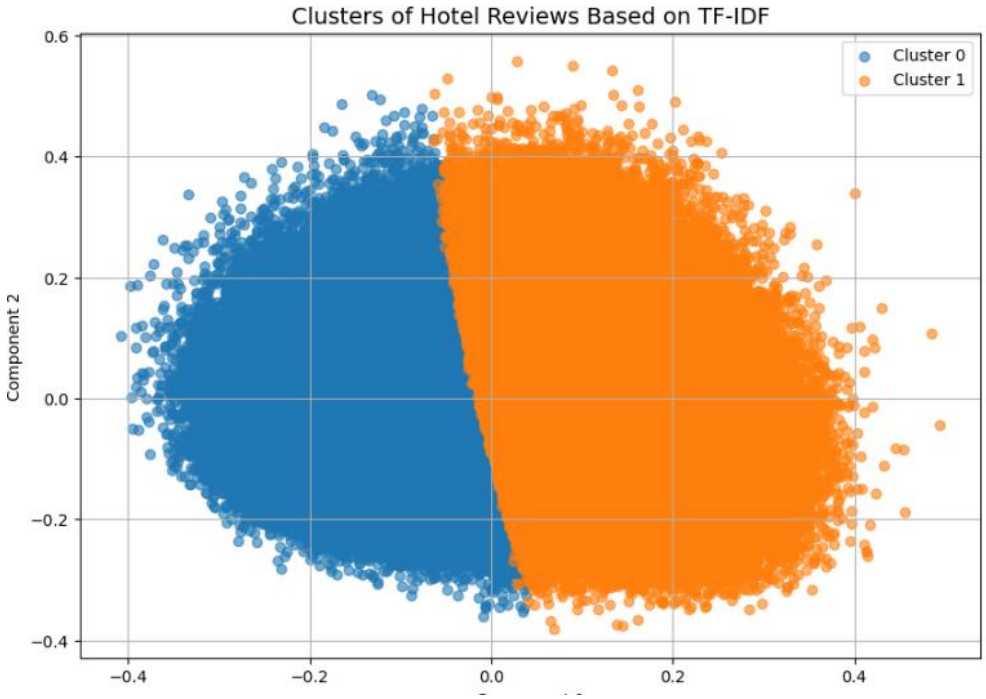


Figure 4.12 The proportion of clustering

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.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

The study aims to identify the pattern of hotel reviews. The hotel reviews dataset is collected from Kaggle, , where reviews are initially stored in a JSON format and later converted into a more manageable CSV format. The data undergoes preprocessing, which includes cleaning (removing punctuation, stop words, and converting to lowercase), tokenization, and ensuring data quality by checking for missing values. Exploratory Data Analysis (EDA) is conducted using Python to uncover patterns and relationships, it can be concluded that the word count and character count have a linear pattern. This indicates the increase in the word count influences the increase in the character count. Apart from that, the correlation between the word count and character count is stronger than the correlation between word count and sentence count. This is because after the dataset goes through the pre-processing phase. Usually, all the punctuation will be removed and it becomes one sentence.

The process of feature extraction utilizes TF-IDF to quantitatively signify the significance of words, focusing on the 500 most commonly used terms. Meanwhile, K-Means clustering sorts the reviews into categories of positive and negative sentiments. The project highlights the necessity of preprocessing steps such as the removal of stopwords and data cleaning to streamline the data for efficient sentiment analysis. The findings indicate a distinct clustering of sentiments, with some areas of overlap and outliers, showcasing the diversity in the content of the reviews. This thorough approach can ensure the precise classification of customer feedback, employing sophisticated methods in data processing and machine learning.

5.2 Future Works

Few gaps have been identified as a result of this research, and these could be addressed in the future. Therefore, this study proposes a few suggestions and ideas that may be useful for potential researchers to further expand research on sentiment analysis on hotel reviews. Additionally, there are a few suggestions for what the authority can do to improve both major and minor sectors in the future. For example, after data pre-processing phase, some data is also noisy. The suggestion is to do an advanced text preprocessing techniques, such as stemming, lemmatization, and part-of-speech tagging to improve the quality of data input for machine learning models.

Second, this study only has two sentiments which are positive and negative. Many reviews may not be entirely positive or negative but instead reflect mixed or balanced opinions. Therefore, expand the sentiment classification from binary which is positive and negative to multi-class analysis such as positive, negative and neutral. This method provides a deeper insight into the dynamics of customer feelings, enabling companies to better comprehend and address specific customer needs. Though this method may present difficulties, like accurately categorizing unclear reviews and navigating overlapping sentiment types, investigating pre-trained models like BERT can improve effectiveness.

Finally, the research employs a conventional method, K-means clustering for sentiment analysis. Therefore, utilizing advanced deep learning models like Long Short-Term Memory (LSTM) or Bidirectional Encoder Representations from Transformers (BERT) can better capture the contextual significance compared to traditional techniques such as Random Forest or K-Means. This particular type of recurrent neural network is tailored to manage sequential data by retaining essential information across lengthy sentences while filtering out irrelevant details. This quality makes it especially proficient in examining the connections between words in a review and addressing complex sentiments, such as mixed feelings expressed within a single remark.