SENTIMENT ANALYSIS OF HOTEL REVIEWS BASED ON NAÏVE BAYES CLASSIFIER

TEOH XU HAO, CHOW HUI TING, PANG WOON YEN

Department of Information Systems, Faculty of Computer Science & Information Technology, Universiti Malaya, Kuala Lumpur 50603, Malaysia

Abstract

The Internet is now a widely available global technology for people for information and knowledge sharing. Reviews have become important for the tourism industry as the number of hotel reviews is rapidly increasing. Compared to hotel reviews, quantitative hotel ratings are not preferable as they do not contain as much information as qualitative text reviews. Due to the large amount of textual data, manual analysations are impractical and laborious. Therefore, sentiment analysis, a natural language processing technique, is implemented to classify the reviews and distinguish them into positive or negative reviews. This paper focuses on the sentiment analysis of London-based hotels' reviews with Naïve Bayes Classifier based on the features extracted using TF-IDF after pre-processing the dataset. The models trained with Naïve Bayes Classifier are also compared with that of Support Vector Machine. With sentiment analysis, time and energy can be conserved as the overall best classifier, Multinomial Bayes Classifier can achieve 92% accuracy in 0.06 computation time. The limitation of this work is that the dataset only consists of single language English text reviews. Besides that, the sentiment analysis only classifies between 2 outcomes, positive or negative, instead of three, which are positive, neutral or negative.

Keywords: Sentiment Analysis, Machine Learning, Naïve Bayes Classification, Support Vector Machine, Hotel Reviews

1.0 Introduction

In this modern era of technology, the Internet has become an informative place where various and distinct types of knowledge are shared. People all around the globe are able to connect and share their views and opinions towards everything, such as interesting places to visit and recommended hotels. All sorts of reviews have been increasing rapidly throughout the years. One of the significantly impacted areas is the tourism industry (Hu et al., 2017). It has become a norm for travellers to survey and study the reviews beforehand to have an excellent experience for their vacation. Since the reviews on the

Internet have become a key and influential factor for travellers, the hotel businesses heavily rely on the reviews to make good impressions. Convincing reviews are recommendations, which is remarkably beneficial and profitable to the hotel owners. Information overload has become a massive predicament as there can be up to thousands of reviews for a single hotel. Suppose important knowledge and insights can be extracted from the reviews. In that case, it is a win-win situation for both parties, as both hotel management and travellers can save their time on reading and analysing the vast amount of long text reviews one after another.

The other information that can be easily utilised is the hotel rating system. Provided by most hotel booking websites, both hotel ratings and reviews are provided to the existing customers to rate and comment. Hotel ratings, for instance, 5 out of 5, are quantitative data. They are simpler to be analysed as the mean can be easily calculated by dividing the total accumulated ratings. However, hotel rating systems are different from written reviews. As customers usually rate a hotel with their overall impression of the hotel, a hotel can be rated 3 out of 5 with an excellent written review (Farisi et al., 2019). Besides that, hotel reviews are usually more detailed than hotel ratings, as they contain more information. The reviewers are able to share more of their thoughts and experience on various aspects like the service, the environment and the location of the hotels. The written reviews matter more as different people have different preferences. Some people might prefer a cosy environment rather than a strategically located hotel. Therefore, the hotel reviews are more valuable than the hotel rating system, as more information can be collected from the data.

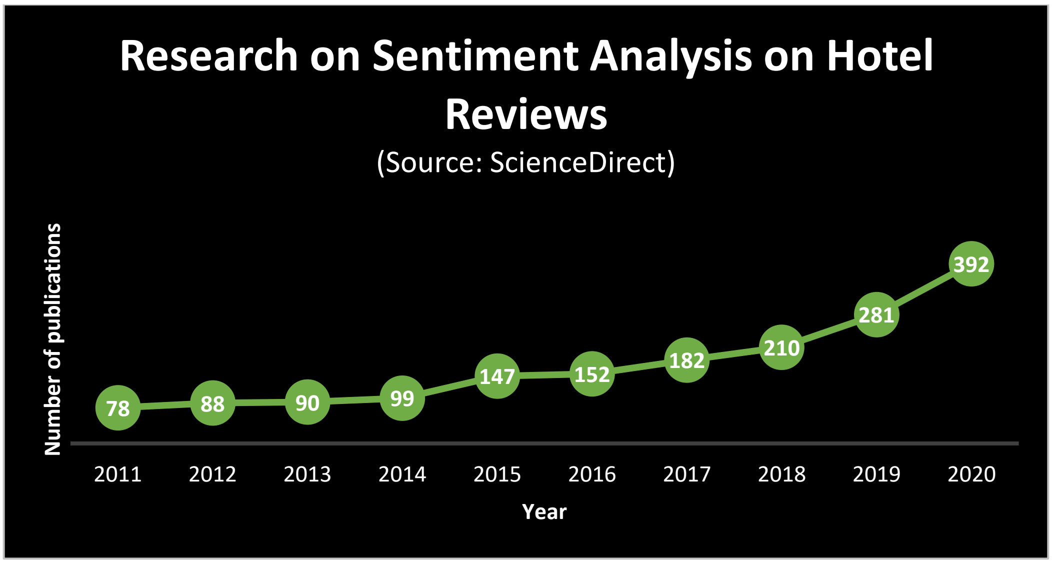
Due to the large scale of textual data, manually analyse and interpreting the reviews to classify them are inconsistent, inefficient and impractical. It is strenuous and laborious to perform repetitive tasks manually and successively. More time and effort should be spent on other resources, such as marketing and strategy, to improve the marketability of products. Therefore, sentiment analysis is introduced. Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique to determine whether data is positive, neutral, or negative. Often performed in qualitative textual data, sentiment analysis is practical and handy for enterprises to monitor and identify the feelings, thoughts and ideas from the consumers' response and reviews. Sentiment analysis is implemented in various domains, such as finance, business and politics (D'Andrea et al., 2015). In addition, sentiment analysis is also applied in social media, restaurant (al Ajrawi et al., 2021)and laptop reviews (Granmo, 2021). The obtained information from the raw textual data is valuable and advantageous to modify the currently existing products and even innovate new products to enhance the competitiveness of organisations in the market. By analysing the emotions behind the reviews of customers, the hotels can build on their quality and reputation based on the customers' demand, which boosts sales and profits and forges customer loyalty.

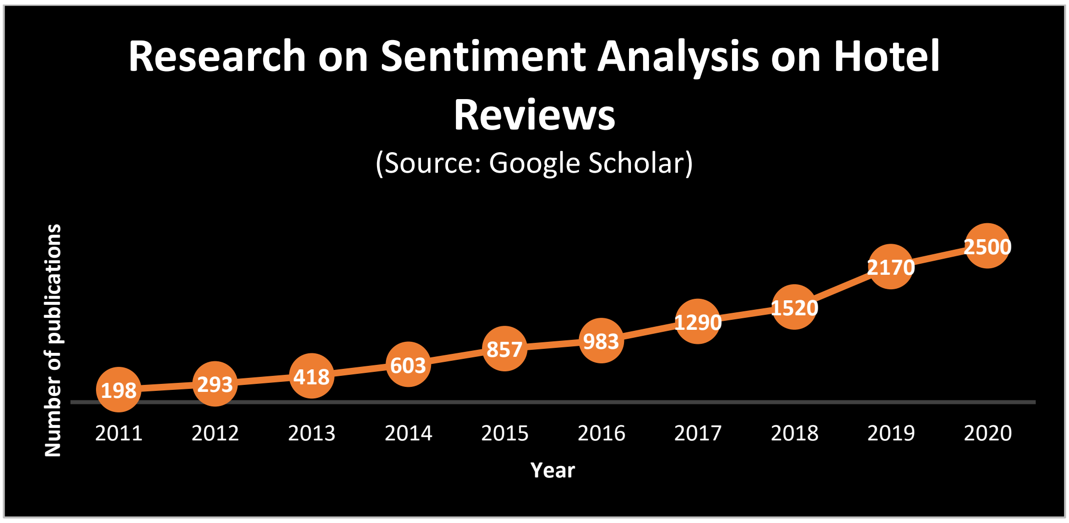
In this research, we are focusing on the sentiment analysis of London-based hotels' reviews. The pre-processing of the Kaggle dataset involves procedures such as case folding, punctuation removal, stop word removal, lemmatisation and tokenisation. After pre-processing, Bag of words model with Term Frequency — Inverse Data Frequency (TF-IDF) score are implemented in feature extraction. Based on the extracted features, Naïve Bayes Classifier and Support Vector Machine are adapted to predict the probability of the hotel review belongs to which sentiment class. The rest of the paper has been planned and arranged as follows. Section 2 provides a literature review about the previous related research done on this topic and concise summaries. This is followed by Section 3, where the research methodology is discussed. Section 4 describes and illustrates the findings and results of our research. In Section 5, the results are analysed and discussed. Finally, the conclusion of the work is presented in Section 6.

2.0 Literature Review

This section is segregated into four parts. The first is the studies that are related to sentiment analysis in general. The second part is the pre-processing of text, followed by feature extraction and selection. Lastly, the classification methods of sentiment analysis on hotel reviews.

The literature search for sentiment analysis on hotel reviews has been conducted on ScienceDirect and Google Scholar. The graphs below illustrated the results of publications when we searched the keywords "Sentiment Analysis on Hotel Reviews". (Data retrieved on 13 May 2021.)





Both graphs show an increasing trend over the past 10 years, from 2011 to 2020. The number of publications on the Google Scholar database is relatively higher as compared to the ScienceDirect database. Papers related to sentiment analysis on hotel reviews are helpful to explore the text pre-processing, features selection and classification methods of sentiment analysis on our topic. The searching process has also been narrowing down to "text mining in hotel reviews" and "text pre-processing in hotel reviews".

The literature review for each part is summarised as follows:

Table 2.1 Summary of Literature Review on General Sentiment Analysis

Study	Concept (Research Motivation)	Field of Study	Remarks

(Andrea et al., 2015)	General approaches that are applicable to opinion mining	Multiple fields	Discussed how sentiment analysis can be applied in finance, business and politics
(Al Ajrawi et al., 2020)	Prediction of ratings by using multiple techniques integrated based on text content	Restaurant Review	Combined sentiment analysis with other techniques (Random Forest) to predict rating based on the text by customers
(Ravi & Ravi, 2016)	Multiple applications and techniques to be applied in sentiment analysis	Social Media	Discussed on the general approaches, tasks and sub-tasks in sentiment analysis. Applications and techniques used as well.
(Yadav et al., 2021)	Aspect-based sentiment analysis to be simplified	Restaurants and laptops reviews	Simplified the pre-processing of aspect-based sentiment analysis.
(Jagtap & Pawar, 2013)	Sentiment analysis on sentence level	Classification of sentiment analysis	Discussed on the differences of different applications, approaches and analysis methods of sentiment analysis.

Table 2.2 Summary of Literature Review on Text Pre-processing

Study	Concept (Research Motivation)	Text Pre-processing Techniques				Remarks
		Casefolding	**Tokenisation**	**Stopword Removal**	**Stemming**	

(Campos et al., 2019)	Limitation of words in text mining	✓	✗	✓	✓	Involved cleaning ununiformed text and created domain to cater text
(Annisa et al., 2019)	Sentiment analysis on hotel reviews using LDA	✓	✓	✓	✓	Utilised the most common text pre-processing techniques
(Fernando et al., 2019)	Double embeddings and attention mechanism on hotel review terms extraction	✓	✓	✗	✓	Involved normalisation to clean the typos or informal text
(Muhammad et al., 2020)	Opinion Mining applying Word2vec and Long Short-Term Memory (LSTM)	✓	✓	✓	✓	Involved padding to append token <pad> until the maximum capacity of the document is reached.

Table 2.3 Summary of Literature Review on Feature Extraction and Feature Selection

Study	**Concept (Research Motivation)**	**Feature Extraction**	**Feature Selection**	**Remarks**

(Akhtar et al., 2017)	Summarise the aspects of hotel reviews	Split the predefined aspects for hotel reviews into 9 categories		
(Farisi et al., 2019)	Multinomial Naïve Bayes classifier on hotel reviews	Bag of words to describe the occurrence of words	Select features based on frequency and remove features with minimal difference between positive and negative classes	
(Yu, 2016)	Aspect-based analysis on hotel reviews	Bag of words, TD-IDF		TD-IDF used to deal with less common but interesting aspects

Table 2.4 Summary of Literature Review on Sentiment Classification

Study	**Concept**	**Classification Method**	**Remarks**
(Fiarni et al., 2016)	Study Naïve Bayes algorithm	NB Classifiers	NB is rather easy to adapt
(Sutabri et al., 2018)	Study NB algorithm with Corpus	MNB Classifier + Corpus Based	Including Corpus based data helps improve accuracy

(Hemalatha & Ramathmika, 2019)	Proposed to calculate the mode of multiple supervised classifiers	NB, MNB, BNB, Logistic Regression, Linear SVC,	Taking the mode of all classification results does not lead to the highest accuracy
(Kurniawan et al., 2018)	Proposed hierarchical classification method	RTF + MNB TF-IDF + MNB	RTF+MNB leads to higher accuracy
(Putri et al., 2019)	Compare SVM and ME	SVM, ME	SVM gives better result
(Srivats Athindran et al., 2018)	Proposed hybrid approach	Lexicon based + NB	Accuracy and robustness ensured
(Kaur et al., 2018)	Proposed new approach	N-gram + KNN (proposed), SVM	Proposed method outperforms SVM
(Rahman & Hossen, 2019)	Compare various supervised models	MNB, BNB, SVM, ME, DT	MNB has the highest accuracy
(Tariyal et al., 2020)	Compare various approaches	Simple linear classifier (LDA), Nonlinear classifiers (CART, KNN), Complex nonlinear classifiers (SVM, RF, C5.0)	CART outperforms other methods
(Zahoor & Rohilla, 2020)	Compare various classifiers	NB, SVM, Random Forest, Long Short-Term Memory	NB has the highest accuracy

(Dey & Noor, 2019)	Compare classification methods with regression methods	NB, DT, RF, KNN, LR, Adaptive Boosting	Classification algorithms have better results
(Singh et al., 2019)	Compare MNB with Bernoulli	NB, MNB, BNB	MNB slightly better than Bernoulli NB

2.1 Sentiment Analysis

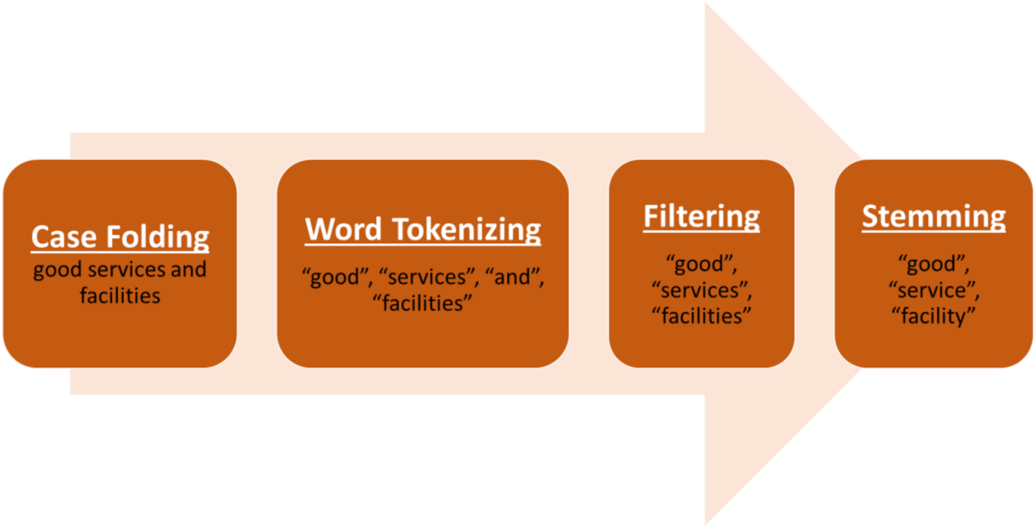
Sentiment analysis is a natural language processing (NLP) method used to decide or determine whether a collected data is positive, negative or neutral. It is also known as opinion mining. Companies often use it to analyse customer reviews and customer feedback to help monitor the product or service sentiment. Sentiment analysis is applied to textual data to analyse people's sentiments or attitudes towards certain entities. Instead of creating questionnaires to collect feedback or ratings from customers, there is plenty of ready info online, for instance, tweets on social media. People share opinions online. Companies can utilise the text over social media platforms to analyse customer behaviours and preferences, used for brand monitoring and market research if the company wants to make a new release product. *(Andrea et al., 2015)* The features of sentiment analysis can also be combined with other techniques such as random forest to form a rating prediction model. which predicts the rating from the text submitted by customers. *(Al Ajrawi et al.,2020)* Sentiment analysis is beneficial, especially in the review analysis for better business strategies and problem-solving by understanding the customer needs.

Application of sentiment analysis in different levels can be made – in document level, sentence level or aspect level. *(Ravi & Ravi, 2015)* Opinion mining that is done at the document level aims to determine the polarity of the document as a whole disregard of any specific aspects. *(Yadav et al., 2021)* When each sentence is examined as an individual unit, it is called sentence-level sentiment analysis. Two tasks are involved for sentiment analysis at sentence-level - classification of subjectivity and classification of sentiment. The aspect-level classification consists of three tasks – identify the feature in the content, assign the polarity value to the feature class, and link the feature to the group feature synonym. *(Jagtap & Pawar, 2013)*

2.2 Text Pre-processing

There are multiple steps to be taken before starting the analysis, which is under the pre-process of text. According to *Campos et al. (2019)*, we should clean the text by converting all the text to lowercase for better comparison with the model created afterwards. Then, remove noise, uninformative text such as advertisements and punctuations. After the basic cleaning of the dataset, there are *three* steps in the pre-process – stemming, removing stop words and utilisation of created domain. Stemming is to reduce a word to the base word, Porter-Stemmer a popular English rule-based stemmer can be used. For example, "swimming" and "swimmer" will be reduced to "swim". Stop words removal aims to remove the typical listed articles and prepositions, such as "in", "of" etc. Next, they created two domains – "Hotel Domain" and "Adjectives" to cater for the words that should be analysed in the sentiment analysis. Eliminate empty reviews after restricted the text with the two domains. In the final step, text transformation is used to give more weight to terms that appear relatively less frequent in the document.

Text pre-processing is the preliminary step before data processing. *(Annisa et al., 2019)* Text pre-processing aims to transform the raw text from human input to an analysable format by machines. The study conducted by *Annisa et al. (2019)* also emphasised that the first step should be case folding (convert all text to lowercase), word tokenising (parsing the text into each token or word), filtering (stop word removal) and stemming (reduce the word to its base form). The text pre-processing for hotel reviews could be illustrated in the process flow below by using "Good services and facilities" as raw text:


Case Folding
good services and facilities
Word Tokenizing
“good”, “services”, “and”, “facilities”
Filtering
“good”, “services”, “facilities”
Stemming
“good”, “service”, “facility”

A research paper conducted by *Fernando et al. (2019)* on Indonesian hotel reviews with aspect-based sentiment analysis started the text pre-processing with the normalisation of sentences. Normalisation process, also known as stemming aims to clean the informal text, abbreviations and typos in customer reviews. Case folding can be carried out simultaneously with normalisation to convert a sentence to lowercase and at the same time, clean up the typos and standardise the words to be formal words. Then, the review is tokenised into a token list by treating a word as a token.

There is one final step in the pre-processing mentioned by *Muhammad et al. (2020)* – padding. Padding is used to append the document with the token "<pad>" until the maximum limit of a document is reached. In most cases, a document in a dataset has fewer words than the maximum limit allowed. In a nutshell, the sub-processes of text pre-processing include case folding, tokenisation, stop word removal, stemming and padding. It is unnecessary to apply all techniques in the text pre-process for one study. We have to adjust them according to our needs. The most common techniques are case folding, tokenisation, stop word removal, and stemming used by many research studies on sentiment analysis of hotel reviews.

2.3 Feature Extraction and Feature Selection

Feature extraction is classifying the reviews into predefined categories in the domain. These categories can be manually defined according to the frequently recurring aspects in the dataset. According to *Akhtar et al. (2017)*, the commonly predefined aspects for hotel reviews are categorised into nine categories: value, location, service, meal, facility, room, quality, staff, and surrounding. The adjectives for each category also be included in the predefined aspects. Bag of words is another way to do feature extraction with text data. It represents text in a frequency table format that describes the occurrence of each word in a document. *(Farisi et al., 2019)* Another technique called TD-IDF (term-frequency multiply inverse of document-frequency) is applied to increase the weightage of less common aspects in the document. *(Yu, 2016)*

Feature selection is an essential process to get the training dataset to build the model. Feature selection will select some features only from the predefined aspects to simplify the model building process. The advantage of feature selection is improving accuracy by eliminating the noise features. There are two approaches for feature selection. Frequency-based selection selects features that have more occurrences, while second approach is to delete features where the difference of probability values between positive and negative classes is minimum. *(Farisi et al., 2019)* The train and test procedures will be effective if we carry out the feature selection to reduce the less relevant features to be in the dataset for modelling.

2.4 Sentiment Classification

This section will discuss the literature review for the classification stage in sentiment analysis.

Sentiment analysis is an effective method to study and investigate the opinion of people. Businesses in many fields can be benefited by using sentiment analysis to understand their customer's review and predict future trends. Many researchers have done a vast amount of research on enhancing the accuracy of sentiment analysis in various approaches. According to *Jagdale et al. (2016)*, the sentiment analysis classification approach can be categorised into the Supervised Learning Approach and the Unsupervised Learning Approach. The Supervised Learning Approach uses supervised machine learning models to classify the sentiment of text data. Examples of the models include K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Multinomial Naïve Bayes (MNB), Maximum Entropy (ME) and more. Unsupervised Learning Approach usually applied when data are not labelled. This approach includes several methods such as Lexicon based classification, Dictionary-based classification, and Corpus based classification. *(Jagdale et al., 2016)*

Fiarni et al. (2016) 's research stated that high sentiment classification accuracy could be achieved by applying the Naïve Bayes model. They proposed a system that can automate the whole sentiment analysis workflow from data collection to sentiment classification. The study showed that the application of the Naïve Bayes model improved accuracy and shortened the processing time of the proposed system. This study took care of the variety of language used by Indonesian and Indonesian slang by creating subjectivity lexicons that record words and polarity created for the specific word. The limitation of this study was that the classified results were not compared with other supervised machine learning methods.

Sutabri et al. (2018) proposed a new approach to perform sentiment analysis. The proposed approach was the combination of the Multinomial Naïve Bayes model and customised Corpus data. Multinomial Naïve Bayes model predicts the conditional probability of each word and estimates the sentiment class (positive, negative, and neutral) of the word. The Corpus data in this study was created from Indonesia hotels online review and article related to the hotel's hospitality. The classification results of the proposed method were compared with the results of the Multinomial Naïve Bayes model with no Corpus data. The authors concluded that the proposed method outperformed the model that was not combined with Corpus data. This indicates that the use of Corpus able to increase the accuracy of sentiment classification.

Naïve Bayes, Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Linear Support Vector Clustering are supervised machine learning algorithms that are suitable for sentiment analysis. *(Hemalatha & Ramathmika, 2019)* The study of *Hemalatha & Ramathmika (2019)* stated that selecting the mode of classification results from the five mentioned algorithms improves the accuracy of the overall experiment results. The study suggested that using more advanced data pre-processing methods and feature extraction methods to analyse long and descriptive sentences can help increase the quality of classification results. However, in this study, only the basic tokenisation method applied to pre-process the text data. This study provided a detailed description of data collection, toolkit usage and steps to perform to experiment. The author gave adequate information to reproduce the research.

Kurniawanet et al. (2018) 's work showed that hierarchical classification performed better than flat classification. The proposed hierarchical classification method involved two-level classification stages. First, Level 1: classify the collected text data into neutral sentences and opinion sentences. Then Level 2: classify the sentiment of opinion sentences. The results indicated that using the Raw Term Frequency feature extraction method with Multinomial Naïve Bayes classification model in stage Level 1 gave better results than using the TF-IDF feature extraction method. Their study proved that the 10-fold cross-validation method helped to prevent adding bias to the proposed model.

Putri et al. (2019) pointed out that the Maximum Entropy model outperformed the Support Vector Machine model when performing sentiment classification tasks. Among the four kernels of the Support Vector Machine (Linear, Polynomial, RBF, Sigmoid), the Linear kernel gave the highest accuracy. Though, the Support Vector Machine Linear kernel model was slightly less accurate when compared to the

Maximum Entropy model. The confusion matrix used by the authors to evaluate the performance of the two mentioned models. The research also showed that the fishbone diagram was a suitable visualisation technique to explore the causal factor of the classified sentiment.

Lexicon-based methods mainly compare text data against a dictionary that contains pre-classified terms. *(Srivats Athindran et al., 2018)* They proposed a model that combined the Lexicon based method with the Naïve Bayes method. The model will first classify the sentiment of pre-processed text data using a Lexicon based method. Then to increase the output confidence level, the Naïve Bayes method will be applied. The study showed that the proposed model was able to predict the sentiment at high accurateness. The study also further explored the classified data to determine the weight of each feature toward the product.

Kaur et al. (2018) proposed a new sentiment analysis method that combined the N-gram algorithm and K-Nearest Neighbours algorithm. N-gram algorithm was used in the proposed model to extract features from the processed review data. K-Nearest Neighbours algorithm was used to classify the sentiment of the review data. The performance of the proposed model was compared with the Support Vector Machine model on the same dataset. The precision, recall and accuracy of the proposed model were all higher than SVM. The results of the study showed that the proposed model gave better results and was able to achieve 86% of accuracy.

Rahman & Hossen (2019) discussed the performance of multiple machine learning models when classifying the sentiment of the text data. The observed machine learning models were Bernoulli Naïve Bayes, Multinomial Naïve Bayes, Support Vector Machine, Maximum Entropy, and Decision Tree. The experiment results showed that Multinomial Naïve Bayes performed better than other models meanwhile, Maximum Entropy had the worst results. The overall performance of various classification models also discussed in other studies. *Tariyal et al. (2020)* compared the performance of multiple machine learning classifiers. The observed classifiers were simple linear classifiers (LDA), nonlinear classifiers (CART, KNN) and complex nonlinear classifiers (SVM, RF, C5.0). The significant finding of this study was the nonlinear classifiers CART approach outperformed other mentioned classifiers. *Zahoor & Rohilla (2020)* compared several supervised classifiers including Naïve Bayes, Support Vector Machine, Random Forest Classifier, Long Short-Term Memory Networks. Reviews regarding six different fields of events were used as input data. Four of the event review classification results showed that the Naïve Bayes classifier had the highest accuracy. *Dey & Noor (2019)* also studied the performance of multiple algorithms when classifying the review's sentiment. Their work indicated that classification algorithms had better results compared to regression algorithms. *Singh et al. (2019)*'s research showed that Multinomial Naïve Bayes algorithm more accurate than Bernoulli Naïve Bayes algorithm.

In conclusion, much research selected Naïve Bayes algorithms as a classification method. Support Vector Machine is another efficient text analysing tool. This study will use Naïve Bayes algorithms and Support Vector Machine to classify hotel review sentiment after pre-processed and extract feature from the dataset.

3.0 Methodology

3.1 System Overview

The final goal of our project is to develop an automatic classifier in terms of hotel reviews. A dataset of London hotel reviews in English has been used by applying machine learning algorithms as classification in this case study. The results of the analysis will be the classifications of reviews into positive opinion or negative opinion. The pre-processing of text is done before the classification process in the sentiment analysis. There are several steps involved in pre-processing stage – case folding, punctuation and stop words removal, lemmatisation or stemming, and tokenisation. After we get the clean dataset in the analysable format, feature extraction and feature reduction are performed to decide the important features to be involved in the analysis stage to ensure the machine learning model's performance.

The context diagram for the system is illustrated as follow:

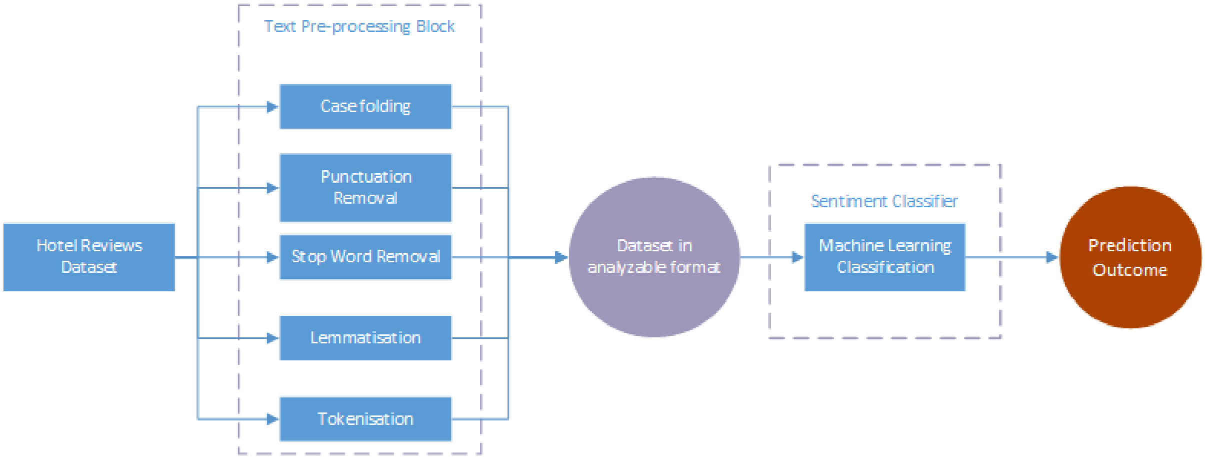


Figure 3.1 Context diagram for the system of sentiment analysis on hotel reviews

3.2 Text Pre-processing

Based on the dataset, there is a column for hotel reviews written in string format. Before entering the data modelling stage, text pre-processing is necessary to transform the data into an analysable format. The Text Pre-processing Block illustrated in Figure3-1 shows the main steps.

The first step is to remove all the html nodes or code-like text within brackets, such as <UTF-8>. The function *BeautifulSoup* from package *bs4* in Python is very useful in this step. Any strings within the brackets will hence be replaced with "". The next step is case folding, which is change all the words in the review column to lowercase. Punctuation removal will be removing all the punctuations like comma, full stop and others. This can be done by using replace function and declaring the regular expression of punctuation. In addition to this, numerical values also have to be removed from the text as the objective of this case study is to predict the sentiment of the text reviews. Thus, numerical values like 930am or 10km do not play a role in the classification. Hence any word with digits will be removed.

Stop words removal can be done by importing the *stopwords* from *nltk.corpus* package. Example of stop words are "in", "the", "just", "a", "what", etc. The frequency of occurrences of stop words will be relatively higher than the number of occurrences of a noun or an adjective, hence eliminating these stop words so that the classification process can capture the important words and make a prediction with better accuracy. Next, lemmatisation is done to modify every word in a sentence to be the base word. The difference between stemming and lemmatisation is that stemming will remove the affixes of a word and does not check the spelling and meaning of the stemmed word. Lemmatisation, on the other hand, will return the base word into the valid English word form. For example, stemming will stem the word "troubling" to be "troubl", but lemmatisation will lemmatise it to be "trouble". Stemming is not preferred in this case study as it does not return a valid English word. Therefore, lemmatisation is processed right after the stop words removal. The last step in text pre-processing is tokenisation. A review contains multiple words, and tokenisation will tokenise the words by splitting every single word. The function of *word_tokenize* from the *nltk* package is handy in this step. Additionally, importing the pos_tag from the nltk package can also add the part-of-speech (POS) tag to every word in the reviews. For example, the word *"stay"* is tagged as *"NN"*, indicating a singular noun. In contrast, the word *"lovely"* is tagged as *"RB"* indicated adverb.

3.3 Feature Extraction

Bag of words model with Term Frequency — Inverse Data Frequency ("TF-IDF") score are implemented in feature extraction. All the unique terms or words contained in the review column is collected in the bag of words, and every review is vectorised to 0s and 1s – if the word exists in that particular review, the value will be 1; otherwise, it will be 0. (Das & Chakraborty, 2018) A simple illustration is shown below:

Pre-processed review 1: good staff service environment fabulous breakfast

Pre-processed review 2: back stay good service

Table 3.1 Sample Bag of Words Model

Review/Words	back	stay	fabulous	breakfast	good	staff	service	environment
Review 1	0	0	1	1	1	1	1	1
Review 2	1	1	0	0	1	0	1	0

Bag of words only presents the existence of a word in a sentence of the review. It does not identify how important the word is in sentiment analysis. TF-IDF is used to identify the important but less frequent words. The calculation is made from two parts, which are TF and IDF. First, the term frequency ("TF") is calculated as the ratio of the number of word occurrences to the total number of words contained in a sentence. The TF can be defined in a mathematical formula as:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{ik}} \tag{1}$$

where

$$i = row\ number$$

$$j = column\ number$$

$$n = number\ of\ occurences$$

$$k = range\ of\ column\ index$$

The inversed data frequency ("IDF") is used to increase the weightage of the rare words in a document. The formula of IDF score can be defined as:

$$idf(word) = \log\left(\frac{N}{f_t}\right) \tag{2}$$

where

$$N = number\ of\ rows\ or\ reviews\ in\ a\ document$$

$$f_t = number\ of\ occurences\ of\ a\ term\ in\ the\ document$$

The TF-IDF score for a particular word can be calculated as:

$$w_{ij} = tf_{ij} \times \log\left(\frac{N}{f_t}\right) \tag{3}$$

A simple illustration of the TF-IDF calculation can be done as follow by using the same example:

Table 3.2 Sample of TF-IDF Calculation

Word	TF		IDF	TF*IDF	
	Review 1	Review 2		Review 1	Review 2
back	0	1/4	log(2/1) = 0.3	0	0.075
stay	0	1/4	log(2/1) = 0.3	0	0.075
fabulous	1/6	0	log(2/1) = 0.3	0.05	0
breakfast	1/6	0	log(2/1) = 0.3	0.05	0
good	1/6	1/4	log(2/2) = 0	0	0
staff	1/6	0	log(2/1) = 0.3	0.05	0
service	1/6	1/4	log(2/2) = 0	0	0
environment	1/6	0	log(2/1) = 0.3	0.05	0

Based on Table 3.2, we can see that the common words have 0 TF-IDF score, which implies that they are not significant. However, the words that only appear once in the document have a TF-IDF score of 0.05 or 0.075. They indicated they carry more weight in the document, which is rare important words.

3.4 Data Visualisation

The popular data visualisation chart for text analysing is a word cloud. A word cloud is to show the density of words in a document. The word with the highest density (highest number of occurrences) will be displayed with the biggest size in the picture.

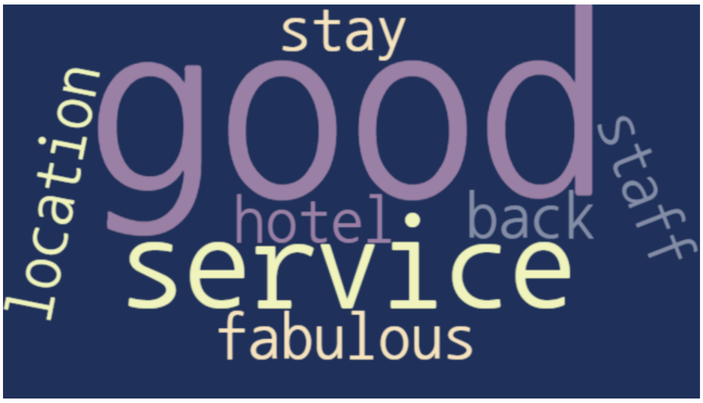


Figure 3.2 Sample Word Cloud

(Generated from wordclouds.com)

Based on Figure 3.2, it is observed that "good" occurred the most in the document, followed by "service". The larger the word, the higher the density.

Heatmap is also a good visualisation in natural language processing. The colour represents a group or classification. The bigger the area of a single colour, the higher is the frequency of that group in the dataset. It can be used to examine the classification and prediction results and check the frequency by groups for single column or multiple columns.

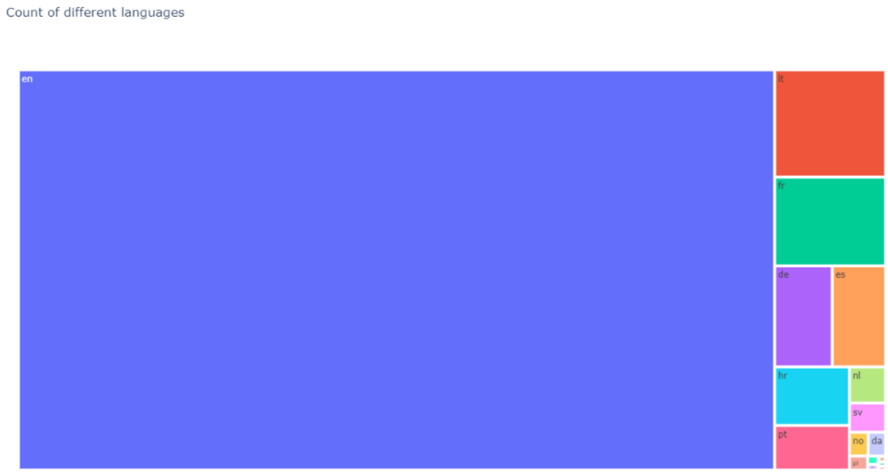


Figure 3.3 Sample Heatmap

3.5 Naïve Bayes

The Naïve Bayes Classifier was derived from the mathematical theorem, the Bayes' theorem. The main purpose of the Naïve Bayes Classifier is to calculate the best-fitted classification within a problem domain for a given piece of data. (Yang, 2018) The upside of using the Naïve Bayes Classifier is that it does not require a large dataset to train the model. On the other hand, it is relatively easy to implement and is very efficient in classifying large text data. (Mamtesh & Mehla, 2019)

Naïve Bayes Classifier refers to a collection of classification algorithms including Multinomial Naïve Bayes, Bernoulli Naïve Bayes, and Gaussian Naïve Bayes. The Multinomial Naïve Bayes is one of the most used algorithms in Natural Language Processing. It assumes multinomial distribution for all pairs and keeps a count of the frequency of the word. The Bernoulli Naïve Bayes is useful when classifying discrete data and is in binary form. The Gaussian Naïve Bayes is used for classifying continuous data.

The Naïve Bayes Classifier assumes that the features are independent of each other. The occurrence of a feature in a particular class does not relate to any other feature. (Mamtesh & Mehla, 2019). These classifiers have been developed to classify the probability of class membership by calculating the probabilistic size of each feature for each class. The algorithm (4) calculates the probability of class j based on the predictor $x_{i.}$ The Bayes' theorem computes the posterior probability P (c_j | x_i) from the probabilities P (c_j), P (x_i | c_j), and P (x_i). The algorithm is shown in (4):

$$P\left(c_j \mid x_i\right) = \frac{P\left(c_j\right) \times P\left(x_i \mid c_j\right)}{P\left(x_i\right)} \tag{4}$$

In (4),

- P (c_j | x_i): Posterior probability, the class j when given word i has occurred.
- P (x_i | c_j): Conditional probability, the occurrence of predictor xi given class j probability.
- P (c_j): Class Prior Priority, the probability of the occurrence of class j.
- P (x_i): The probability of occurrence of the word i.

3.6 Support Vector Machine

Support Vector Machine is another robust text classifier. This machine learning model can be useful for classification, outlier detection, and Regression. Compared to Naive

Bayes Classifiers, Support Vector Machine requires a lesser training dataset to train the model. However, it requires more computation time to tune the model. The Support Vector Machine is effective when the number of dimensions is higher than the sample. This classifier draws a hyperplane line to classify the data distributed in different classes. The hyperplane separates the space into two classes. One contains the features that belong to a particular group, and the other contains features that do not belong to that group. The Support Vector Machine provides different functions, which are also called the kernel. The kernels are used to determine the hyperplane. This study makes use of the kernel, which separates the maximum possible distance between two classes.

3.7 Proposed Study

This study adapts Naïve Bayes Classifiers and Support Vector Machine to predict the probability of the hotel review belongs to which sentiment class based on the features extracted using TF-IDF after pre-process the dataset. The hotel review dataset is captured via Kaggle.com. Python 3 is the main tool used for doing the sentiment analysis tasks. Python 3 provides packages such as nltk and sklearn that are useful for analysing text data. The basic workflow of the methodology is described in Figure 3.4.

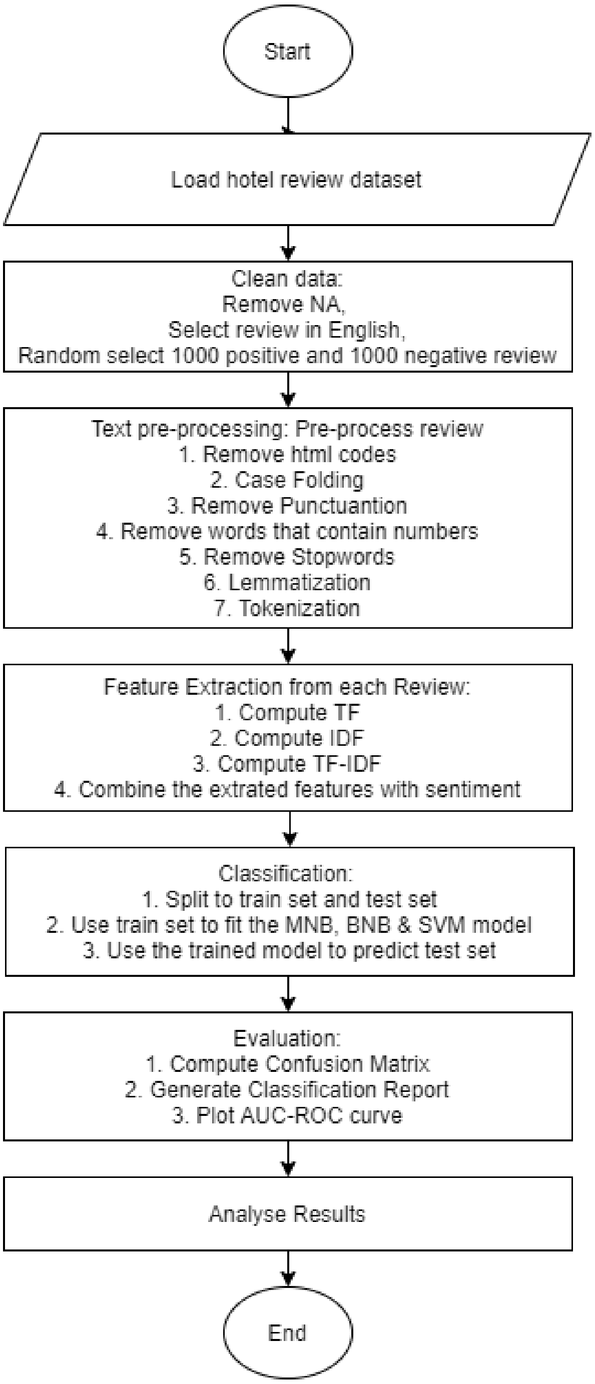

Start
Load hotel review dataset
Clean data:
Remove NA,
Select review in English,
Random select 1000 positive and 1000 negative review
Text pre-processing: Pre-process review
1. Remove html codes
2. Case Folding
3. Remove Punctuantion
4. Remove words that contain numbers
5. Remove Stopwords
6. Lemmatization
7. Tokenization
Feature Extraction from each Review:
1. Compute TF
2. Compute IDF
3. Compute TF-IDF
4. Combine the extrated features with sentiment
Classification:
1. Split to train set and test set
2. Use train set to fit the MNB, BNB & SVM model
3. Use the trained model to predict test set
Evaluation:
1. Compute Confusion Matrix
2. Generate Classification Report
3. Plot AUC-ROC curve
Analyse Results
End

Figure 3.4: Basic Workflow of this Study

4.0 Results

4.1 Dataset

Reviews of London-based hotels has been utilised as the dataset in this study. The dataset is obtained from Kaggle, where the dataset is built by PromptCloud using the in-house web-scraping method. There are 6 attributes in the dataset: Property Name, Review Rating, Review Title, Review Text, Location of the Reviewer and Date of Review.

Property Name is the name of the hotel which are based in London. Review Rating rank from 1 to 5 where 1 indicates very bad and 5 indicates very good. Review Title and Review Text are in string format, and the text is not entirely in English. Location of the Reviewer contains the city name or the country name of the reviewers, ranging from the United States, the United Kingdom, to Asia Pacific countries. The date of review ranges from the year 2003 – 2018.

There are 27,331 tuples in the original dataset. This case study is only performed on English text reviews; hence a sample of English reviews is drawn. The train test data is drawn by selecting the balanced dataset between positive reviews and negative reviews to prevent introducing bias to the model afterwards. There will be 1000 rows of positive reviews and 903 rows of negative reviews. Neutral reviews are omitted in this study as our objective is to predict sentiment classification, which is either positive or negative.

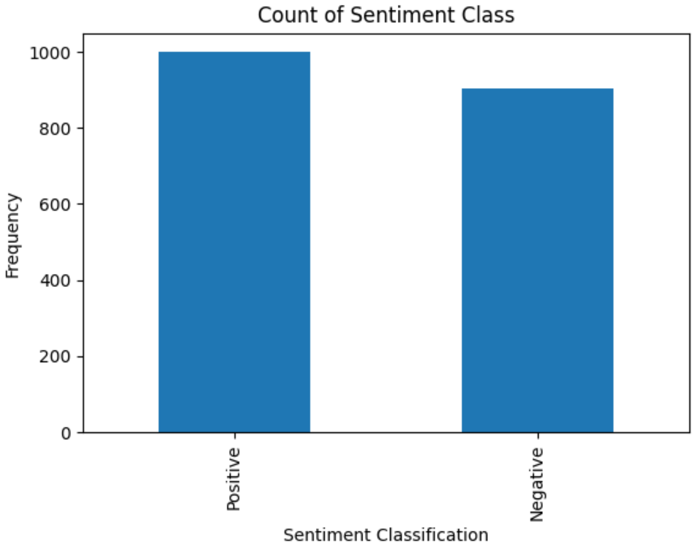


Figure 4.1 Frequency of Sentiment Class for Sample Drawn

Figure 4.1 shows the frequency of positive reviews and negative reviews in the sample drawn.

4.2 Results of Text Pre-Processing

After went through all the text pre-processing steps, the tokenised text can be visualised by checking the most common words in the sample drawn.

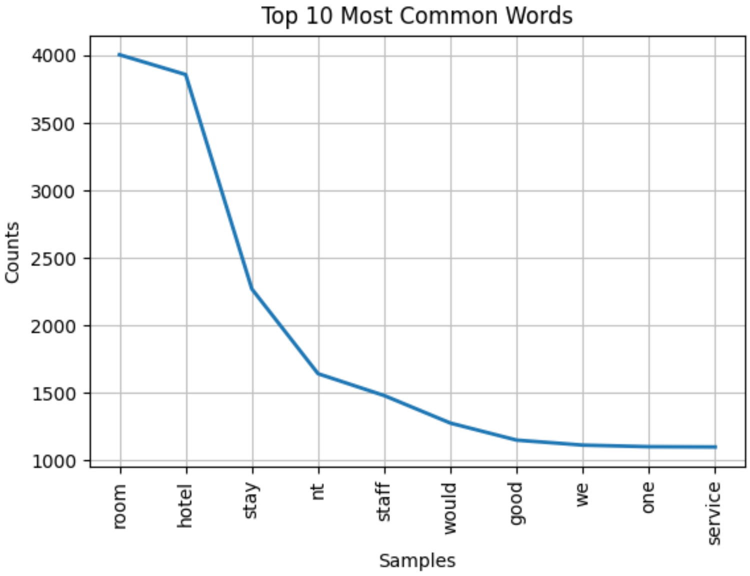


Figure 4.2 Top 10 Most Common Words in the Sample

Figure 4.2 displays the top 10 most common words that we are having in the review column in the sample drawn. Alternatively, a word cloud can be used to illustrate the density of each word in the text reviews as well. However, we could not tell the overall sentiment based on these two charts.

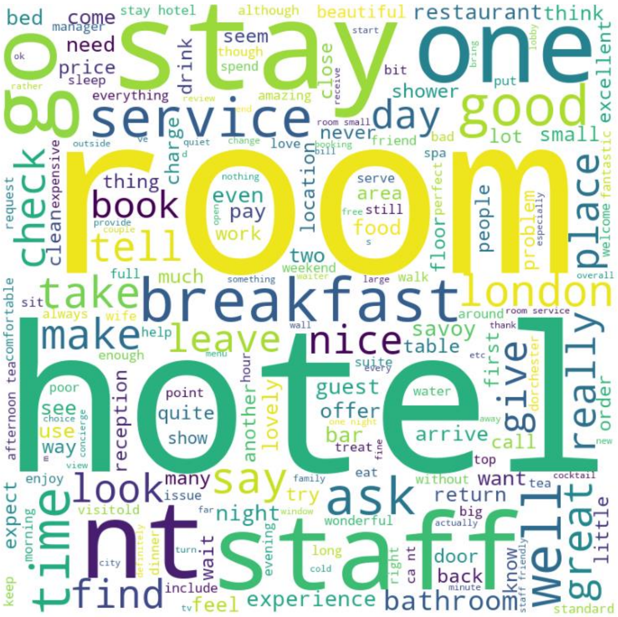


Figure 4.3 Word Cloud of Tokenized Text

4.3 Results of Classifiers

The performance of classifiers is evaluated using a Confusion matrix. The confusion matrix is often used for evaluating the machine learning model. Table 4.1 below shows how to interpret the confusion matrix:

Table 4.1: Confusion Matrix

	Predicted: Negative	Predicted: Positive
Actual: Negative	TN	FP
Actual: Positive	FN	TP

- TN: True Negative, the negative review classified as negative.
- TP: True Positive, the positive review classified as positive.
- FN: False Negative, the positive review classified as negative.
- FP: False Positive, the negative review classified as positive.

Accuracy, precision, recall, and F-score are calculated for each classifier. The algorithms below show how the criteria calculated:

$$Accuracy = \frac{TP \times TN}{TP + TN + FP + FN} \tag{5}$$

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{7}$$

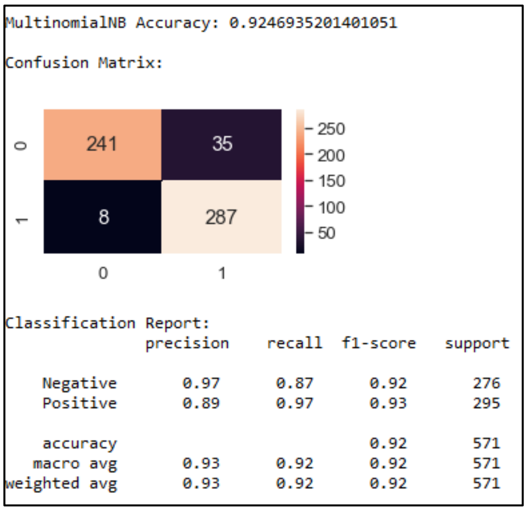
$$Recall\ (P) = \frac{TP}{TP + FN} \tag{8}$$

$$Precision\ (N) = \frac{TN}{TN + FN} \tag{9}$$

$$Recall\ (N) = \frac{TN}{TN + FP} \tag{10}$$

In (5), the accuracy is calculated to show how often the classifier is correct. F-score in (6) is calculated to find the weighted average of the recall and precision. In (7), the precision is how often the classifier is correct when predicting the 'Positive'. In (8), the recall ratio is calculated to represent how often the classifier predicts 'Positive' when it is actually 'Positive'. In (9), calculates the ratio of the total number of reviews correctly classified as the class 'Negative'. In (10), the recall ratio is calculated to represent how often the classifier predicts 'Negative' when the review data is actually 'Negative. The F-score for class 'Negative' can be calculated using (6).

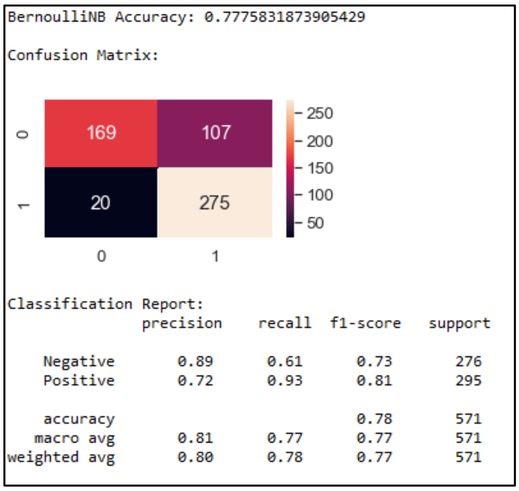
Figures 4.4 to Figure 4.6 show the performance of Multinomial Naïve Bayes Classifier, Bernoulli Naïve Bayes Classifier, and Support Vector Machine. The dataset is separated into 1331 training data and 571 test data. The training data contains 626 negative reviews and 705 positive reviews. The test data contains 276 negative reviews and 295 positive reviews. The performance of classifiers when classifying test data was evaluated. The negative class is represented as 0 and the positive class is represented as 1 in the confusion matrix.



```
MultinomialNB Accuracy: 0.9246935201401051

Confusion Matrix:
```

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.97      0.87      0.92       276
    Positive       0.89      0.97      0.93       295

    accuracy                           0.92       571
   macro avg       0.93      0.92      0.92       571
weighted avg       0.93      0.92      0.92       571
```

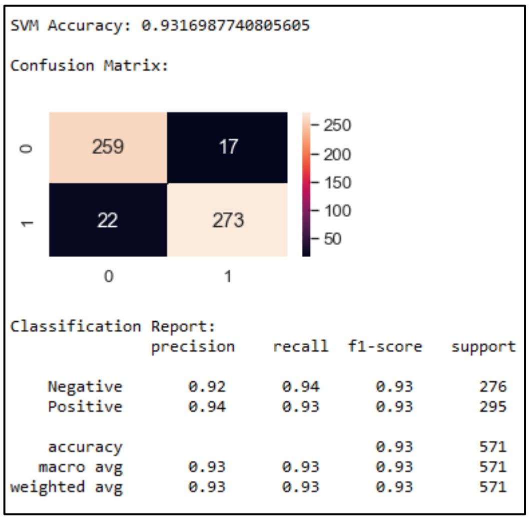
Figure 4.4: Results of Multinomial NB Classifier



```
BernoulliNB Accuracy: 0.7775831873905429

Confusion Matrix:
```

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.89      0.61      0.73       276
    Positive       0.72      0.93      0.81       295

    accuracy                           0.78       571
   macro avg       0.81      0.77      0.77       571
weighted avg       0.80      0.78      0.77       571
```

Figure 4.5: Results of Bernoulli NB Classifier



```
SVM Accuracy: 0.9316987740805605

Confusion Matrix:
```

```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.92      0.94      0.93       276
    Positive       0.94      0.93      0.93       295

    accuracy                           0.93       571
   macro avg       0.93      0.93      0.93       571
weighted avg       0.93      0.93      0.93       571
```

Figure 4.6: Results of SVM Classifier

5.0 Discussion

The top 10 most common words extracted from the sample are room, hotel, stay, nt, staff, would, good, we, one and service. The performance evaluation of classifiers is summarised in Table 5.1. The time taken for each classifier to train the model is plotted in Figure 5.1. The Support Vector Machine has the highest accuracy, 93% and the Bernoulli Naive Bayes has the lowest accuracy, which is 77%. However, the Support Vector Machine takes the longest time to train the model, which is 1.69 seconds. The Multinomial Naive Bayes takes only 0.006 seconds to train the model and is the shortest time taken. Hence, the Multinomial Naive Bayes has the best performance since it requires the shortest time to process and has 92% accuracy.

Table 5.1: Performance Comparison of Classifiers

Classifier	Accuracy	Precision		Recall	
		Negative	Positive	Negative	Positive
Multinomial NB	92%	97%	89%	87%	97%
Bernoulli NB	77%	89%	72%	61%	93%
SVM	93%	92%	94%	94%	93%

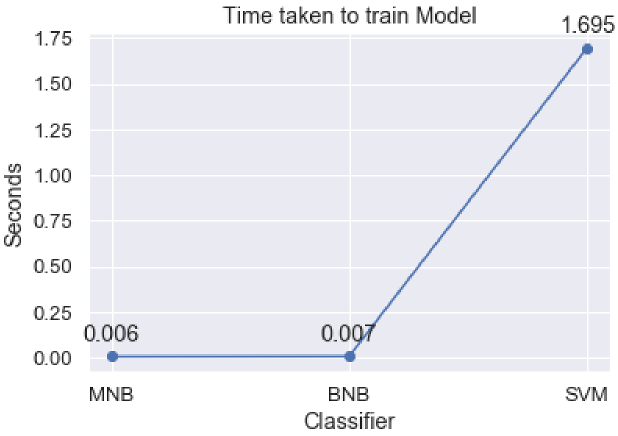


Figure 5.1: Time taken to train Model

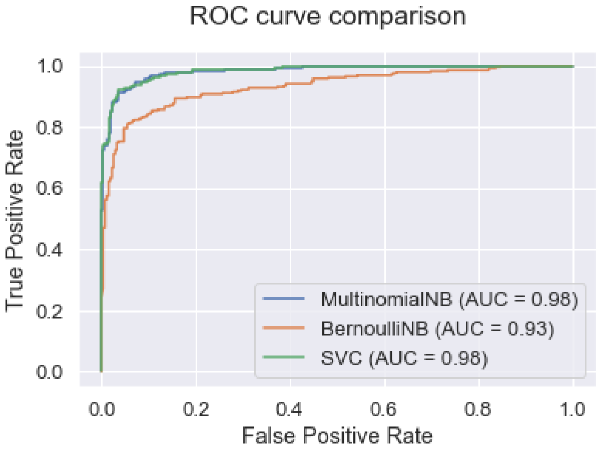


Figure 5.2: ROC curve comparison

The Receiver Operator Characteristic (ROC) curve is plotted to reveal the performance of the classifiers when classifying two sentiment classes. ROC plots the True Positive rate against the False Positive rate with a different threshold. Area Under the

Curve (AUC) of each classifier represents its ability to classify the binary classes. Figure 5.2 shows the AUC-ROC curve of the compared classifiers.

The Multinomial Naive Bayes and Support Vector Machine have identical ROC curve and AUC. Their AUC is 0.98, and the AUC of Bernoulli Naive Bayes is 0.93, which are greater than 0.5 and less than 1. This indicates that the three of the classifiers perform well when classifying between two classes. The classifiers can distinguish positive reviews from negative reviews.

6.0 Conclusion

Sentiment analysis is a practical technique to analyse hotel reviews with high accuracy and short computational time. In this study, Naïve Bayes Classifiers, which are Multinomial Naïve Bayes and Bernoulli Naïve Bayes are implemented. Compared to the Support Vector Machine, which has the highest accuracy of 93%, the Multinomial Naïve Bayes model has achieved a closely similar accuracy of 92%, whereas the Bernoulli Naïve Bayes falls short with 77% accuracy. Nevertheless, both the Naïve Bayes Classifier models have shorter computation time than that of Support Vector Machine, which is 0.006 seconds for Multinomial Naïve Bayes, 0.007 seconds for Bernoulli Bayes Classifier and 1.695 seconds for Support Vector Machine. As for the Receiver Operator Characteristic (ROC) and Area Under the Curve (AUC), The Multinomial Naive Bayes and Support Vector Machine have the same ROC and AUC. The AUC of Multinomial Naive Bayes and Support Vector Machine is 0.98, whereas the AUC of Bernoulli Naive Bayes is 0.93. With their values more than 0.5, it is indicated that all three classifiers performed well in classification. All classifiers are able to distinguish between positive and negative reviews. Considering both accuracy and time taken for model training as important parameters, Multinomial Naïve Bayes is the best model as it requires 92% of accuracy, only 1% lower than the Support Vector Machine, with the shortest training time of 0.006 seconds, which is crucial as real-life applications will have larger training datasets. The limitation of this work is that our training dataset only consists of single language English text reviews. Besides that, the sentiment analysis only classifies between 2 outcomes, positive or negative, instead of three, which are positive, neutral or negative. The extension of this work can be on the sentiment analysis of hotel reviews with multiple languages and different algorithms with a dataset that consists of positive, neutral and negative reviews.

7.0 References

1. Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based Sentiment Oriented Summarisation of Hotel Reviews. *Procedia Computer Science,* 563 - 571. doi:10.1016/j.procs.2017.09.115
2. Al Ajrawi, S., Agrawal, A., Mangal, H., Putluri, K., Reid, B., Hanna, G., & Sarkar, M. (2020). Evaluating business Yelp's star ratings using sentiment analysis. *Materials Today: Proceedings,* -. doi:10.1016/j.matpr.2020.12.137
3. Andrea, A. D. ', Ferri, F., & Grifoni, P. (2015). Approaches, Tools and Applications for Sentiment Analysis Implementation. *International Journal of Computer Applications,* 26 - 33. doi:10.5120/ijca2015905866
4. Annisa, R., Surjandari, I. & Zulkarnain (2019). Opinion Mining on Mandalika Hotel Reviews Using Latent Dirichlet Allocation. *Procedia Computer Science,* 161, 739 - 746. doi:10.1016/j.procs.2019.11.178
5. Campos, D., Silva, R. R., & Bernardino, J. (2019). Text Mining in Hotel Reviews: Impact of Words Restriction in Text. *SCITEPRESS,* 442 - 449. doi: 10.5220/0008346904420449
6. Das, B., & Chakraborty, S. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation. Retrieved from http://arxiv.org/abs/1806.06407
7. Dey, U. K., & Noor, A. (2019). For Sentiment Analysis Using NLP. 2019 International Conference on *Computer Communication and Informatics (ICCCI),* 1–6. Retrieved from https://ieeexplore.ieee.org/abstract/document/8821801
8. Farisi, A. A., Sibaroni, Y., & Faraby, S. A. (2019). Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier. *Journal of Physics: Conf.,* 012024. doi:10.1088/1742-6596/1192/1/012024
9. Fernando, J., Khodra, M. L., & Septiandri, A. A. (2019). Aspect and Opinion Terms Extraction Using Double Embeddings and Attention Mechanism for Indonesian Hotel Reviews. *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA),* 1 - 6. doi:10.1109/ICAICTA.2019.8904124
10. Fiarni, C., Maharani, H., & Pratama, R. (2016). Sentiment analysis system for Indonesia online retail shop review using hierarchy Naive Bayes technique. *2016 4th International Conference on Information and Communication Technology, ICoICT 2016, 4(c).* https://doi.org/10.1109/ICoICT.2016.7571912

11. Hemalatha, S., & Ramathmika, R. (2019). Sentiment analysis of yelp reviews by machine learning. *2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019, (Iciccs),* 700–704. https://doi.org/10.1109/ICCS45141.2019.9065812

12. Hu, Y. H., Chen, Y. L., & Chou, H. L. (2017). Opinion mining from online hotel reviews – A text summarisation approach. Information Processing and Management, 53(2), 436–449. https://doi.org/10.1016/j.ipm.2016.12.002

13. Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2016). Sentiment Analysis of Events from Twitter Using Open Source Tool. *International Journal of Computer Science and Mobile Computing, 5(4),* 475–485. Retrieved from http://www.cs.cornell.edu/People/pabo/movie-review-data

14. Jagtap, V., & Pawar, K. V. (2013). Analysis of different approaches to Sentence-Level Sentiment Classification. *International Journal of Scientific Engineering and Technology*, 164 - 170.

15. Kaur, S., Sikka, G., & Awasthi, L. K. (2018). Sentiment Analysis Approach Based on N-gram and KNN Classifier. *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications,* 13–16. https://doi.org/10.1109/ICSCCC.2018.8703350

16. Kurniawan, S., Kusumaningrum, R., & Timu, M. E. (2018). Hierarchical Sentence Sentiment Analysis of Hotel Reviews Using the Naïve Bayes Classifier. *2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018*, 104–108. https://doi.org/10.1109/ICICOS.2018.8621748

17. Mamtesh, & Mehla, S. (2019). Sentiment Analysis of Movie Reviews using Machine Learning Classifiers. International Journal of Computer Applications, 182(50), 25–28. https://doi.org/10.5120/ijca2019918756

18. Muhammad, P. F., Kusumaningrum, R., & Wibowo, Adi. (2020). Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews. *Procedia Computer Science,* 728 - 735. doi:10.1016/j.procs.2021.01.061

19. Putri, B. A. D., Khasanah, A. U., & Azzam, A. (2019). Sentiment analysis on grab user reviews using support vector machine and maximum entropy methods. *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 468–473. https://doi.org/10.1109/ICOIACT46704.2019.8938527

20. Rahman, A., & Hossen, M. S. (2019). Sentiment Analysis on Movie Review Data Using Machine Learning Approach. *2019 International Conference on Bangla Speech and Language Processing, ICBSLP 2019,* 27–28. https://doi.org/10.1109/ICBSLP47725.2019.201470

21. Ravi, K., & Ravi, V., (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems,* 14 - 46. doi:10.1016/j.knosys.2015.06.015

22. Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019,* 593–596. https://doi.org/10.1109/ICACTM.2019.8776800

23. Srivats Athindran, N., Manikandaraj, S., & Kamaleshwar, R. (2018). Comparative Analysis of Customer Sentiments on Competing Brands using Hybrid Model Approach. *Proceedings of the 3rd International Conference on Inventive Computation Technologies, ICICT 2018,* 348–353. https://doi.org/10.1109/ICICT43934.2018.9034283

24. Sutabri, T., Suryatno, A., Setiadi, D., & Negara, E. S. (2018). Improving naïve bayes in sentiment analysis for hotel industry in Indonesia. *Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018.* https://doi.org/10.1109/IAC.2018.8780444

25. Tariyal, A., Goyal, S., & Tantububay, N. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express, 6(4),* 300–305. https://doi.org/10.1016/j.icte.2020.04.003

26. Yadav, R. K., Jiao, L., Goodwin, M., & Granmo, O. (2021). Positionless aspect based sentiment analysis using attention mechanism. *Knowledge-Based Systems,* 1 - 12. doi:10.1016/j.knosys.2021.107136

27. Yang, F. J. (2018). An implementation of naive bayes classifier. Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018, 301–306. https://doi.org/10.1109/CSCI46756.2018.00065

28. Yu, Y. (2016). Aspect-based Sentiment Analysis on Hotel Reviews. *Arxiv Prepr,* 10.

29. Zahoor, S., & Rohilla, R. (2020). Twitter Sentiment Analysis Using Machine Learning Algorithms: A Case Study. *Proceedings - 2020 International Conference*

on Advances in Computing, Communication and Materials, ICACCM 2020, 194–199. https://doi.org/10.1109/ICACCM50413.2020.9213011