UNIVERSITY OF ECONOMICS AND LAW

COURSE: DATA ANALYSIS USING R/PYTHON

Sentiment analysis for hotel reviews using Deep Learning

GROUP 8

List of members:

K204111813 Tran Thi Kieu Trinh (Team leader)

K204110566 Tran Nguyen Ngoc Hieu

K204111797 Nguyen Thi Linh Anh

K204110605 Nguyen Thanh Phat

K204110611 Bang Nhan Tri

Supervisor:

Nguyen Phat Dat, M.A.

Ho Chi Minh City, 2023

Source code: link

ABSTRACT

With the growth of the internet and online sites, a large number of reviews have been left on those platforms, which easily leads to a number of results in terms of influencing hotels and user preferences. As a result, exploring public reviews can provide us with helpful information by understanding users' opinions and has a variety of potential uses. However, the obstacles faced in natural language processing (NLP) keep hampering the effective and accurate polarity detection of sentiment analysis. Many approaches have recently been proposed to obtain insights from these data but there are lots of challenges in dealing with text of huge size. In some recent studies, it has been demonstrated that deep learning models are a promising solution to the challenges of NLP. This study relied on the most research findings which used deep learning to analyze sentiment analysis problems, including such sentiment polarity and then conduct a comparison between 3 models: LSTM, Under-sampling LSTM and DistilBert. Utilizing a method includes data crawling from TripAdvisor, preprocessing, transform, train model and then evaluating and testing on Flask App. The result indicates that DistilBert tends to give the best performance with Accuracy up to 88% as well as other measures of Precision, Recall and F1-Score are 91%, 88% và 89%, respectively. It also reduces bias in favor of a specific polarity, thereby bringing more objective efficiency when analyzing users' perspectives and obtaining meaningful insights.

Keywords: sentiment analysis; hotel reviews; deep learning; LSTM; distilBert

Contents

Figures

Tables

1. Introduction

In order to maintain or enhance the company's financial capacity in volatile markets, the business must re-examine, compile, organize, and improve all aspects related to its business in light of the growing range of sales and increasing competition in the industry. The company must be able to effectively manage its customer relationships to increase revenue and achieve its objectives (Ghazian et al.,2016). Customers can now quickly contact businesses in various ways, one of which is through an online review on a company's website.

It is acknowledged that the most influential form of communication between service providers and consumers, as well as among consumers, is informal communication via internet technology regarding the use of particular products or services and their descriptions (Park et al., 2018). According to Hu and Chen (2016), approximately 89% of global travelers and 64% of global hoteliers are persuaded that online hotel reviews influence hotel bookings. According to the survey that was carried out by Ady and Quadri-Felitti (2015), almost 95% of tourists decided to book a hotel by first reading reviews about it online. Many tourists consider online hotel reviews to be one of the most important factors in their decision to select a hotel. Customers can use online ratings and reviews to make decisions, but reviews provide a better understanding of the hotel (Akhtar, 2017).

Customer reviews of hotel services and products become a factor in the evaluation and improvement of hotel management as well as a trigger for the correction of errors and deficiencies in hotel products or services (Geetha et al., 2017). Reviews composed on the site are huge and unstructured data, so they can be investigated utilizing fitting methods (Gandomi and Haider, 2015). Moreover, there are two main approaches to the sentiment classification process, particularly Lexicon based and Machine Learning based. Additionally, the lexicon-based technique includes estimating orientation for a text based on the semantic orientation of the document's words or phrases. Building classifiers from labeled examples of texts or phrases is the text classification strategy, which is effectively a supervised classification task (Taboada, 2011). Machine learning is often context sensitive and may be coupled with huge language models to provide a reasonably accurate image of the linguistic content of a text. When comparing traditional assessment criteria, machine learning models often outperform lexicon-based algorithms on sentiment analysis tasks (Öhman, 2021). Therefore, machine learning based approach will be used in this study.

Moreover, in this paper, the researchers began by obtaining the term list sentiment per aspect by collecting data on aspect classifications based on similar aspects as training data for sentiment classification by using Long-Short Term Memory (LSTM)

which relies on machine learning based approach. For more details, Long-Short Term Memory (LSTM) is a popular approach. According to Intellipaat (2023), Long-term dependencies can be learned using a variety of Recurrent Neural Networks (RNNs), especially in sequence prediction problems. Because it has feedback connections, LSTM is able to process the entire data sequence, not just single data points like images. Speech recognition, machine translation, and other applications make use of this. LSTM is a unique type of RNN that performs exceptionally well on a wide range of problems. Furthermore, the Long-Short Term Memory (LSTM) has a "Forget Gate" that decides which data should be kept or deleted and refreshes the cell's memory (Do et al., 2019). LSTM sentiment classification was used in the experiment to determine the best performance in sentiment classification.

There are also a variety of papers that have researched about this field and used the same method to analyze the customer's sentiment. Khotimah's study (2019) has proposed some favorable aspects for developing their business and a few negative aspects that the hotel manager should pay attention to and prevent. Unfortunately, the hotel managers need more than that. They would like the technique that can measure and anticipate the effect level whether those aspects are good or not. Additionally, according to Agarwal (2022), the study has proved that sentiment analysis can help the manager identify the risks quickly to solve them immediately. Though the paper has also shown the most impactful social platform that can affect the business performance, the study has not clarified for each specific industry, especially hotel management.

This study aims to to examine customer satisfaction and dissatisfaction via their reviews using deep learning. Several models are proposed before evaluating their performance and determining the most suitable algorithm for this dataset. Finally, several approaches for further research are concluded to enhance analysising performance and contribute more managerial implications.

2. Theoretical background and related work

2.1. Sentiment Analysis

Sentiment analysis is a process of extracting information about an entity and automatically recognizing any of the subjectivities of that entity. The goal is to identify whether user-generated text reflects their positive, negative, or neutral opinions (Dang et al., 2020).

Sentiment analysis is categorized into three levels: document level, sentence level and aspect level (Behdenna et al., 2018; Khan et al., 2016). The first level identifies and analyzes the document classification as a positive, negative or neutral class. The second level makes an analysis of the review sentences, which are then classified as a positive, negative or neutral class. Reviews or comments that express just one emotion are best categorized using this method. Lastly, the aspect level requires its aspect categories to identify and extract features within each sentiment (i.e., product, service or employee) to classify the polarity.

Currently, there are three approaches to address the problem of sentiment analysis: sentiment analysis methods based on sentiment lexicon, sentiment analysis methods based on machine learning, and sentiment analysis methods based on deep learning (Bhavitha et al., 2017).

Lexicon-based techniques were the first to be used for sentiment analysis. They are divided into two approaches: dictionary-based and corpus-based (Salas et al., 2017). In the former type, sentiment classification is performed by using a dictionary of terms, such as those found in SentiWordNet and WordNet. Nevertheless, corpus-based sentiment analysis does not rely on a predefined dictionary but on statistical analysis of the contents of a collection of documents, using techniques based on k-nearest neighbors (k-NN) (Huq et al., 2017), conditional random field (CRF) (Pinto et al., 2003), and hidden Markov models (HMM) (Soni and Sharaff, 2015) among others.

Machine-learning-based techniques (Zhang and Zheng, 2016) proposed for sentiment analysis problems can be divided into two groups: (1) traditional models and (2) deep learning models. Traditional models refer to classical machine learning techniques, such as the naïve Bayes classifier (Parveen et al., 2016), maximum entropy classifier, or support vector machines (SVM). The input to those algorithms includes lexical features, sentiment lexicon-based features, parts of speech, or adjectives and adverbs. The accuracy of these systems depends on which features are chosen. Deep learning models can provide better results than traditional models. Different kinds of deep learning models can be used for sentiment analysis, including CNN, DNN, and RNN.

The hybrid approaches (Pandey et al., 2017) combine lexicon- and machine-learning-based approaches. Sentiment lexicons commonly play a key role within a majority of these strategies. Figure below shows a classification of sentience-based deep-learning techniques.

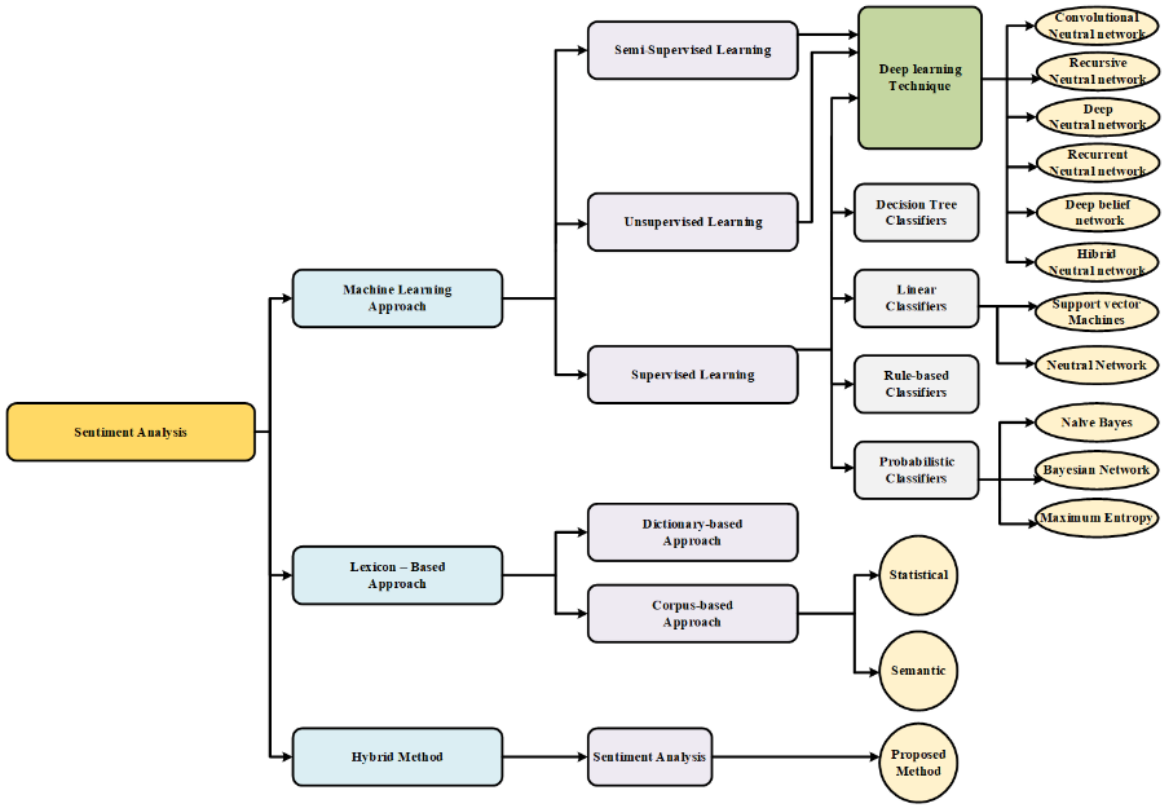

Figure 1: Taxonomy of sentiment analysis techniques (Source: Bhavitha et al., 2017; Medhat et al., 2014).

2.2. Sentiment Analysis for hotel reviews

With the increasing popularity of online platforms in our everyday lives, websites for social networks have become an important and significant source of data on user reviews in a variety of areas. The more reviews that surface help potential customers make product choices to some degree and provide some positive feedback for businesses. For example, if a user wants to choose a comfortable hotel for his vacation, he will browse the review sites to read the views of other users (Shi and Li, 2011). Tourism has expanded how it operates online in the last ten years so in the article, the method is presented that uses the Fuzzy C-means clustering algorithm to autonomously identify sentiment and classify hotel ratings given by customers from one of the major tourism sites. Various methods, such as Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Logistic Regression, and Random Forest, were used to

evaluate hotel reviews. An ensemble learning model combining the five classifiers was also suggested, and the outcomes were compared (Anis et al., 2020).

But in the research of (Ray et al., 2020), using method that took an organized method firstly employs a combination of a binary classification named Bidirectional Encoder Representations from Transformers (BERT) model with three stages for positive -negative, neutral-negative, and neutral-positive emotions combined using a score giving procedure. And then gave these pre-trained word embeddings produced by the BERT models, together with other textual characteristics such as Word2vec word vectors, TF-IDF of common words, subjectivity score, and so on, to a Random Forest classifier. Besides, the Naive Bayes Classifier algorithm also was used in order to make it simple to categorize motels based on emotions analysis. The sentiment analysis process in this study consists of three stages: (1) text pre-processing, which includes transform case, stopword removal, and stemming; (2) the implementation of N-Gram features, namely Unigram, Bigram, Trigram, and the N-Gram feature is a feature that contains a collection of words that will be referred to in the next process; as well as (3) the hotel review process of classification utilizing Menggunakan Bayes Classifier with the search outcomes demonstrating that Unigram is capable of producing superior test results than Bigram and Trigram, having an estimated average accuracy rating of 81.30% (Tiffani, 2020).

2.3. Deep Learning for Hotel Reviews Sentiment Analysis

Sentiment analysis involves the extraction and study of hidden data derived from online platforms in the manner of unstructured data. This study will discuss and propose sentiment analysis approaches using deep learning, a combination of techniques that incorporates both lexicon-based and deep learning methods, and effectively contain business information from unstructured data (Maw and Khine, 2020). In some previous research, online reviews about hotels were analyzed via machine learning algorithms, and an incredibly effective machine learning technique, known as the deep learning method of LSTM, was applied. Furthermore, it is examined how LSTM has surpassed traditional machine learning methods because the results and efficacy of deep learning methods are more successful compared to machine learning algorithms such as Naive Bayesian Classifiers, SVM classifiers, and Decision Trees (Arora and Bansal, 2022). Long short-term memory (LSTM) and gated recurrent units (GRUs) were used for analyzing the hotel review data in the suggested technique, with an accuracy rate of detecting customer opinion of 86% and 84%, correspondingly. Naive Bayes, Decision Tree, Random Forest, and SVM are also used to evaluate the information. The accuracy for Naive Bayes is 75%, the accuracy for Decision Tree is 71%, the accuracy for Random Forest is 82%, and the accuracy for SVM is 71%. Deep learning is used to improve the performance of businesses, acquire feedback from consumers, and forecast

opinions regarding customer feedback (Hossen et al., 2021). The technique of deep learning has already been shown to deliver accurate results, and as it is contrasted to current methods, the outcome in the future is likely to be significantly superior, which will undoubtedly benefit the tourism sector (Arora et al., 2022). When compared with machine learning techniques, deep learning methods create an efficient resolution for various text mining issues such as document clustering, classification, web mining, summarization, and sentiment analysis. The 515 K hotel review collection is used for experimentation. The research obtained superior outcomes utilizing the LSTM model when compared to regular machine learning models and the sequential RNN model, and the experimental findings received additional support by using bidirectional LSTM (BiLSTM) and enhanced LSTM (ILSTM), which increased performance by 2.73% and 5.67%, accordingly (Bhoi et al., 2023).

Many of the studies proposed to date for sentiment analysis use BERT, Sun thorough research into the various BERT text classification fine-tuning techniques led to cutting-edge outcomes for review sentiment analysis (Sun et al., 2019). Additionally, Song established a new benchmark for aspect-based sentiment analysis by investigating the potential of BERT intermediate layers to improve BERT fine-tuning (Song et al., 2020). Alaparthi compared the relative effectiveness of four sentiment analysis techniques and proved the undisputed advantage of BERT in text sentiment classification (Alaparthi et al., 2021). In numerous NLP tasks, BERT has demonstrated its significant text representation advantage. However, due to the complex structure of BERT, a lot of training time and expensive training costs are required (Alaparthi et al., 2021).

2.4. LSTM

Deep learning adapts a multilayer approach to the hidden layers of the neural network. Features are specified and extracted using feature selection techniques or manually in conventional machine learning approaches. Deep learning models, on the other hand, automatically learn and extract features, improving accuracy and efficiency. In general, classifier model hyperparameters are also evaluated automatically. Figure below compares the classification of sentiment polarity using deep learning versus conventional machine learning (Support Vector Machine [SVM], Bayesian networks, or decision trees]. The best solutions are presently offered by artificial neural networks and deep learning in the areas of speech and image recognition as well as natural language processing (Dang et al., 2020).

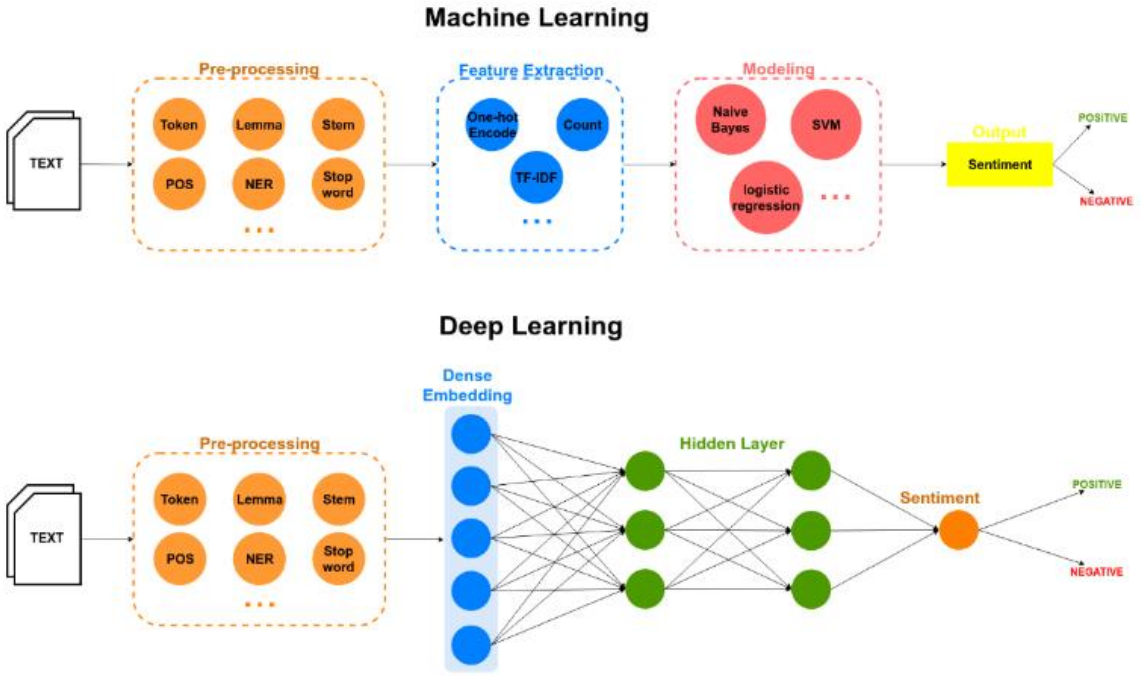

Figure 2: Differences between two classification approaches of sentiment polarity, machine learning (top), and deep learning (bottom). Part of Speech (POS); Named Entity Recognition (NER); Term Frequency-Inverse Document Frequency (TF-IDF). (Source: Dang et al.)

Recurrent neural networks are a subclass of neural networks where the links between neurons are organized into directed cycles, creating feedback loops within the RNN. The handling of sequential data using the internal memory that is recorded by directed cycles is the main function of an RNN. Compared to conventional neural networks, RNNs have the benefit of being able to reuse data from previous computations by applying it to the next input in the sequence. A specific type of RNN called Long Short-Term Memory (LSTM) can use long memory as the input for activation functions in the buried layer. The first time this was stated was by Hochreiter and Schmidhuber (1997).

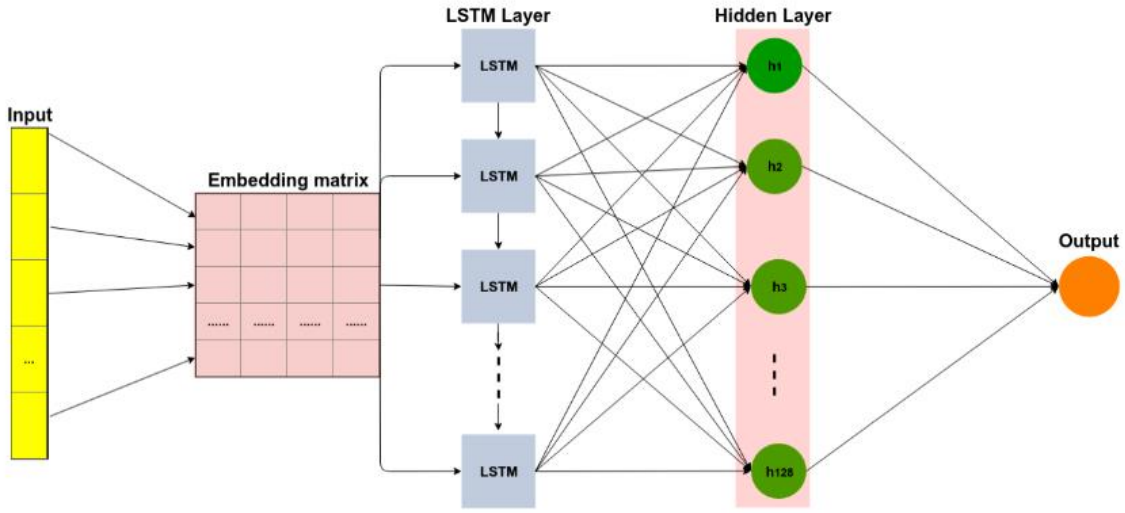

Figure 3: A long short-term memory network. LSTM, long short-term memory. (Dang et al., 2020)

2.5. DistilBert

BERT is the first thoroughly unsupervised, bidirectional language representation model to be created. Before BERT, a number of other bidirectional unsupervised learning-based language models had been taught (Peters et al., 2018). Bidirectional encoder representations from transformers (BERT) is a pre-trained language model that is made to take into account a word's context concurrently from its left and right sides (Devlin et al., 2019).

DistilBERT is created by applying the knowledge distillation method to BERT (Bucila et al., 2006; Hinton et al., 2015), which results in a model that is lighter and faster while maintaining a reasonable level of performance on most of the evaluated tasks. DistilBERT is a smaller general purpose language representation model. DistilBERT shrinks the original BERT model by 40% while maintaining 97% of its language understanding skills and increasing speed by 60%. BERT is the teacher in the process of information distillation, and DistilBERT is the pupil. A small model that represents the student is taught to mimic the behavior of the larger model (i.e. the teacher) (Marco et al., 2022).

2.6. Imbalanced Dataset

Numerous attempts have been made to address the issue of data imbalance in the field of neural networks. The three main streams of the aforementioned solutions to the imbalance issue encompass nearly all of the work. The specific implementations of sampling or cost-sensitive methods, or their combinations, on neural networks are of special interest, though the specifics may vary. Kukar provided several methods for modifying the back-propagation learning technique for multilayered feed-forward

neural networks in a cost-sensitive manner. By incorporating cost factors into various portions of the original back propagation algorithms, he outlined four methods for learning cost-sensitive neural networks. Therefore, cost-sensitive classification, adjusting the network's output, adjusting the learning rate, and minimizing the costs associated with misclassification are suggested (Kukar et al., 1998).

Any data set that shows an unequal distribution between its classes can be considered imbalanced (He et al., 2009). However, the prevailing belief in society is that data sets showing significant, and in some cases extreme, imbalances communicate with imbalanced data. This type of disparity is known as a between-class imbalance, and it is common to see these on the scales of 100:1, 1,000:1, and 10,000:1, where one class significantly out represents another (He & Shen, 2007, Kubat et al., 1998).

The issue with this classification challenge is the use of data intrinsic distinctiveness. The training data's lack of density, the presence of small disjuncts, the identification of noisy data, the dataset shift between the training and test distributions, and the significance of borderline instances are all represented in Figure below and will make it easier to develop the current models in these areas.

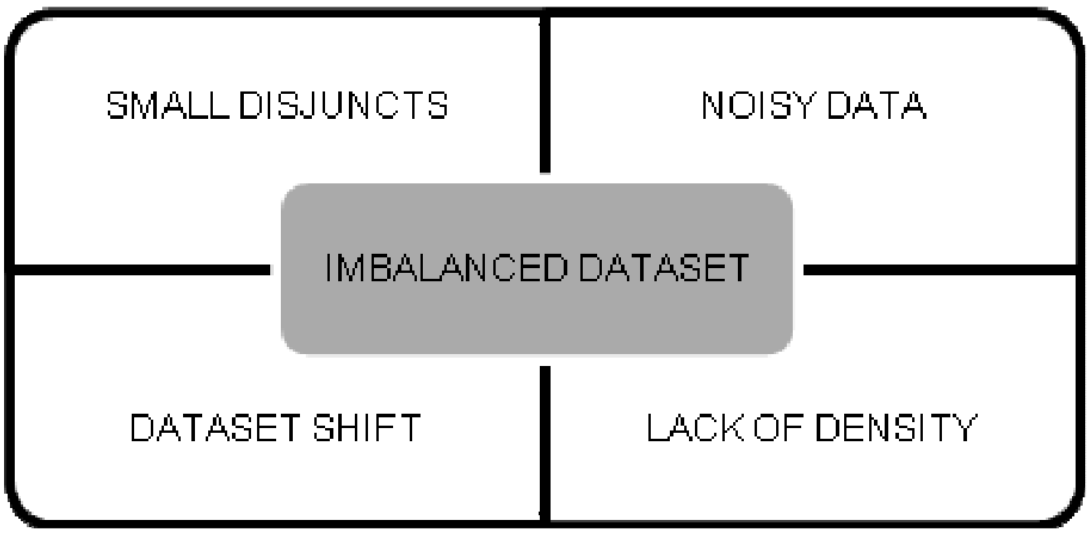

Figure 4: Imbalanced dataset characteristics (Source: Lopez et al., 2013).

3. Methodology and Experiment

In this chapter, the process is illustrated into a picture consisting of several stages. These stages include crawls reviews, preprocessing, transform, train model, evaluate and run flask app. In detail, the data is crawled from Tripadvisor by using The Python library BeautifulSoup. Then, the data is traslated into English, lowercase and removed punctuation, stop words and lemmatized text. Next, the data undergo the EDA before being transformed to prepare for model training. There are 3 different models in the Train Model phase, including LSTM, Under sampled-LSTM and distilBert that is shown below.

3.1. Methodology

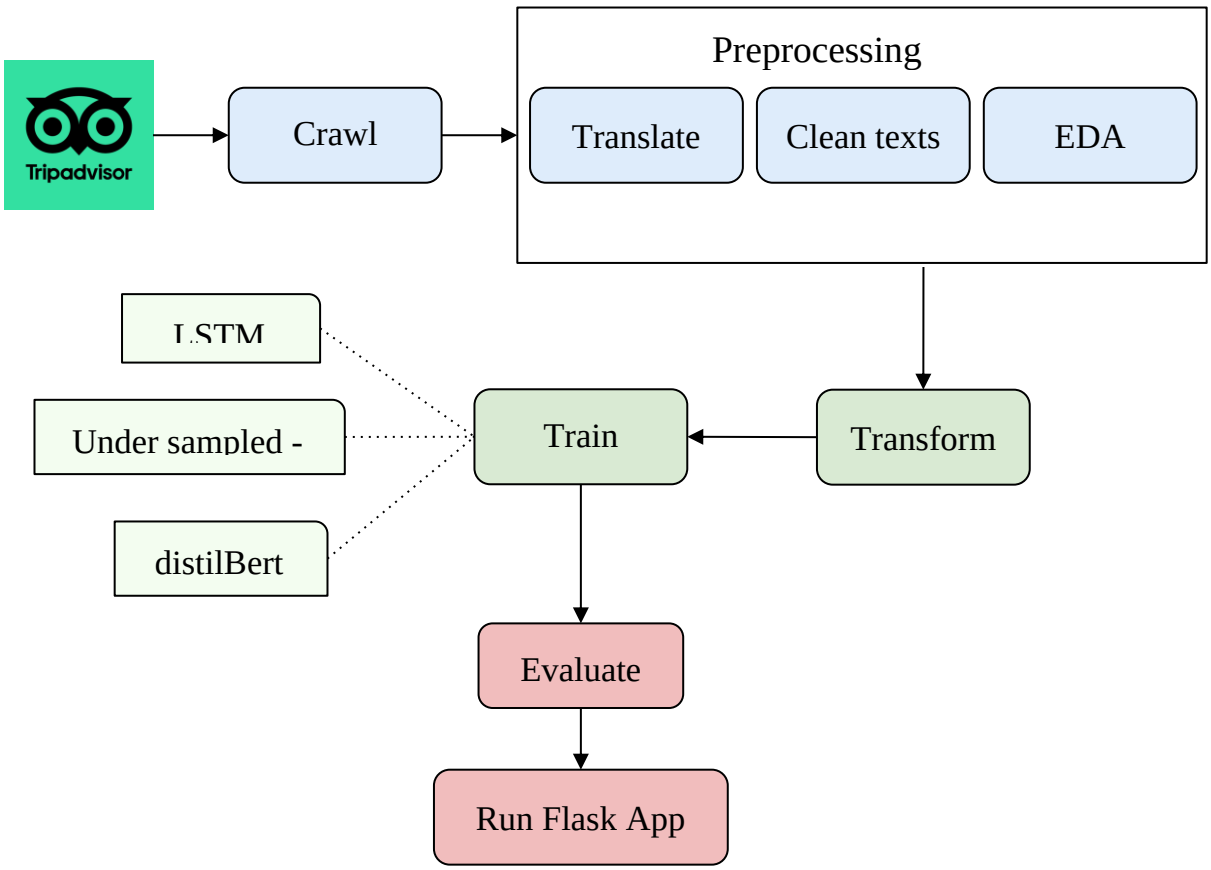

Figure 5: Experimental process (Source: Authors)

First of all, the hotel reviews data is crawled from Tripadvisor and prepared for being trained by going through the "Preprocessing phase". This phase includes: translate all reviews into English, clean texts, and EDA to gain more insights from the data set. Next, the purified data will be transformed for better understanding and forearmed for the following stage. Then, the data will be trained by using these 3 models: LSTM, Under sampled - LSTM and distilBert. After that, the results will be evaluated and prepared for running Flask App.

3.2. Experiment

3.2.1. Data Collection

The Python library BeautifulSoup (Hajba and Hajba, 2018) is used to handle, extract necessary information when accessing the TripAdvisor website's API, and has collected data including reviews, related information about the reviews such as hotel information, user information, and information about the user's actual vacation. The collected dataset consists of 83195 records, including information fields which will be detailed in the following section.

3.2.2. Preprocessing

3.2.2.1. Text cleaning

Lowercase

Customers input content in the customer reviews using both lower- and upper-case letters, which violates grammar conventions. Case sensitivity is a common feature of many of the study's methodologies. As a result, it is challenging for the classifier to determine the orientation of the supplied text. Simply converting the complete text to a standard format could solve this problem. Conversely, the lower (txt) statement is used if we want to carry out the same procedure directly. All upper case writing is converted to lowercase, but all other characters are left alone.

Remove Punctuation

Nearly 40 to 50 percent of the text in a written paper is made up of punctuation. There is no correlation between punctuation and the results of any sentiment analysis algorithm. Since these punctuation marks don't affect the mood analysis, they must be removed. We presented the data in their normalized form at this point and deleted all punctuation from the text. The final content is condensed and simplified. The text was stripped of all punctuation.

Remove Stop Words

Text files often contain phrases that are repeated. Therefore, it is essential to remove the end words. Stop words, in fact, never give the written content any importance. The content frequently contains a lot of words of this nature. As a consequence, text mining has become challenging, and classifier results have been unexpected. At this point, the stop words are removed from the selected data. This approach reduces textual content data while enhancing system effectiveness.

Lemmatize text

Lemmatization is the process of breaking down words into their simplest version, or lemma. lemmatization takes into consideration the context of the word and its

grammar, which makes it more accurate. As it eliminates all tenses from a document and streamlines the text of each one, this normalization method is helpful.

3.2.2.2. EDA (Preprocessing)

The preprocessed dataset has a total of 83,189 records. Details of the collected data fields are shown in Table 1.

Only two columns “Rating” and “Review” from the data set, which are necessary for this study, are kept. The rating field is the user's rating of the hotel on a 1-5 star rating scale. The sentiment of each review is determined based on its rating. To be more specific, reviews having at least 4-star are labeled as positive, while reviews labeled as negative have 1 to 2 star, and 3-star reviews belong to neutral polarity. The sentiment polarities are then encoded as 0 - negative, 1 - neutral, and 2 - positive. Finally, the word_count field is the number of words per user review.

Table 1: Details of the data fields in the preprocessed dataset (Source: Authors)

Field	Describe	Amount	Type
Rating	The rating given by the user for hotel	83,189	int
Review	Contents of the review	83,189	object
Sentiment	Polarity of review	83,189	int
word_count	Word count of the review	83,189	int

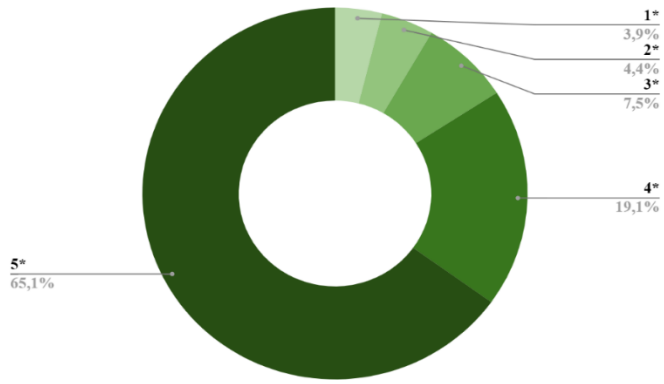

Figure 6: Review allocation according to the rating in the Preprocessed dataset (Source: Authors)

The rating spectrum of the data set is a 1-5 star rating frame, for each type of rating, the dataset has a rating distribution as shown in Figure 6 with most of the ratings. from users is positive (4-5 stars) with a total rate of 84.2% of the total rating of the dataset.

The sentiment score is ranged from 0 to 2, for each type of sentiment, the data set has a distribution of comments according

to sentiment as shown in Figure 7 with most of the reviews from users being positive (sentiment = 2). with 69980 comments.

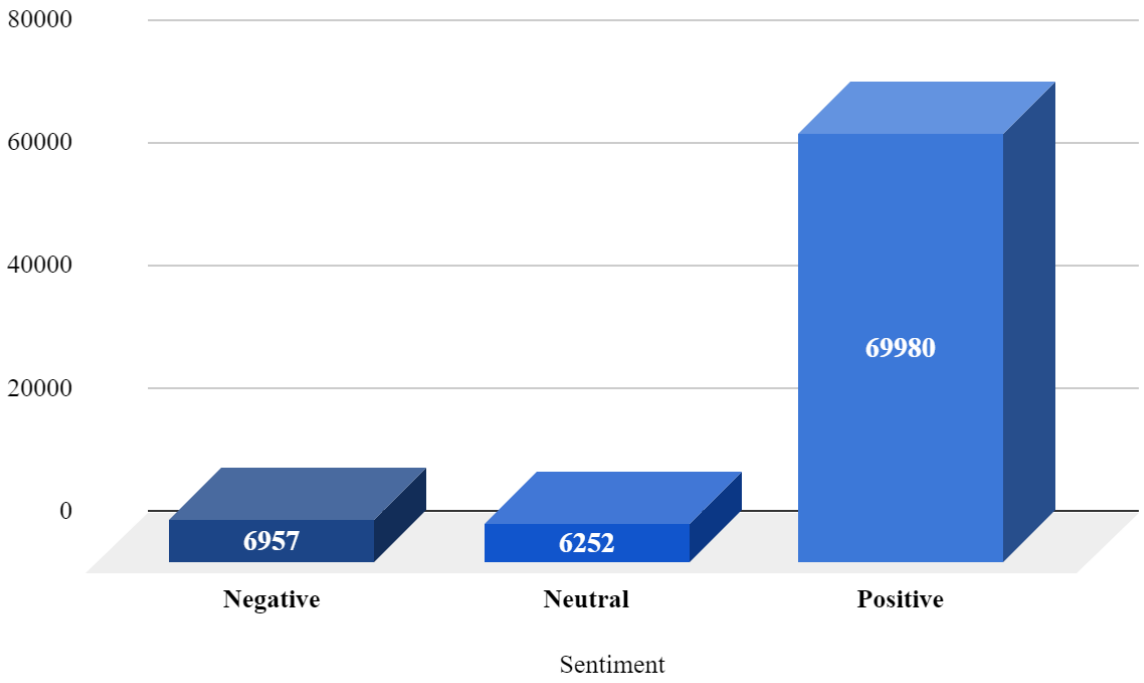

Figure 7: Review distribution according to polarity in the Preprocessed dataset (Source: Authors)

The boxplot chart in Figure 8 shows the distribution of word_count by each sentiment point. Word count allocation in a comment based on sentiment doesn't make too big of a difference:

- For a polarity 0 (negative), the number of words per comment measured in units of measure recorded values from about 5 to 400. The box indicates a value range from about 50 to 200, with a value The median is at the left skewed position. There are some outliers marked as single points scattered from about 380 to 3300. All of these values are outliers and are not within the range allocated in the box.
- For a polarity 1 (neutral), the number of words per comment measured in units of measure records values from about 4 to 390. The box indicates a value range from about 40 to 180, with values The median is at the left skewed position. There are some outliers marked as single points scattered from about 380 to 2500. All these values are outliers and are not within the range of values allocated in the box.
- For a polarity 2 (positive), the number of words per comment measured in units of measure recorded values from about 4 to 290. The box indicates a value range

from about 30 to 130, with a value The median is at the left skewed position. There are some outliers marked as single points scattered from about 250 to 2300. All these values are outliers and are not within the range of values allocated in the box.

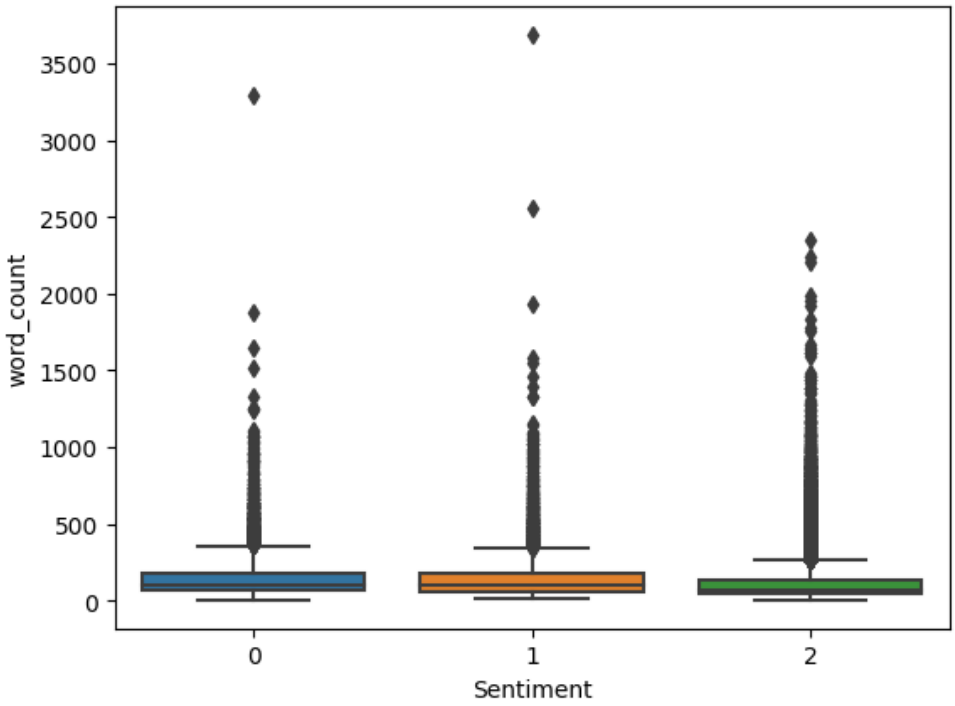

Figure 8: Allocate word count of each review according to polarity (Preprocessed dataset) (Source: Authors)

The word cloud of the entire preprocessed dataset in Figure 9 focuses on words such as: resort, hotel, restaurant, swimming pool, staff. This reflects that users tend to be interested in these categories of hotels.

Figure 9: Word cloud of the Preprocessed dataset (Source: Authors)

Categorized by each sentiment level of the data set, words often mentioned in user comments are shown in Figure 10, 11 and 12. Word Clouds are not too different between sentiment levels when focusing on words like hotel, located, staff,...

Figure 10: Word cloud of negative review (polarity 0)

Figure 11: Word cloud of neutral review (polarity 1)

Figure 12: Word cloud of positive review (polarity 2)

(Source: Authors)

From these word cloud, it can be seen that staff, check-in service and room seem to be the main factors leading to negative experience of customers. Besides, location and sightseeing (such as beach, tourist, etc) seem to meet the customer satisfaction.

Data train

There are 66,551 comments total in the train dataset. Table 2 displays specifics regarding the data categories that were gathered.

The data fields in the train data set have relevance to the study's subject. The ID field makes it easier to separate the comments from one another. On a range of 1 to 5, the rating field requests the user to rate the hotel. The user's sentiment is rated in the sentiment section based on the comments made about the hotel. The sentiment field has a number between 0 and 2, with 0 representing negative emotions, 1 being neutral, and 2 positive emotions. The word count per user comment is displayed in the word_count field.

Table 2: Details of the data fields in the Train dataset (Source: Authors)

Feature	Describe	Amount	Type
ID	ID of the review	66,551	int
Rating	The rating given by the user for hotel	66,551	int
Review	Contents of the review	66,551	object
Sentiment	Polarity of review	66,551	int
word_count	Word count of the review	66,551	int

The train dataset has a rating distribution for each category of rating, as shown in Figure 13, and the rating spectrum is a 1–5 star rating frame. A total of 84% of the ratings for the train dataset are good, with the majority of user evaluations being favorable (4-5 stars).

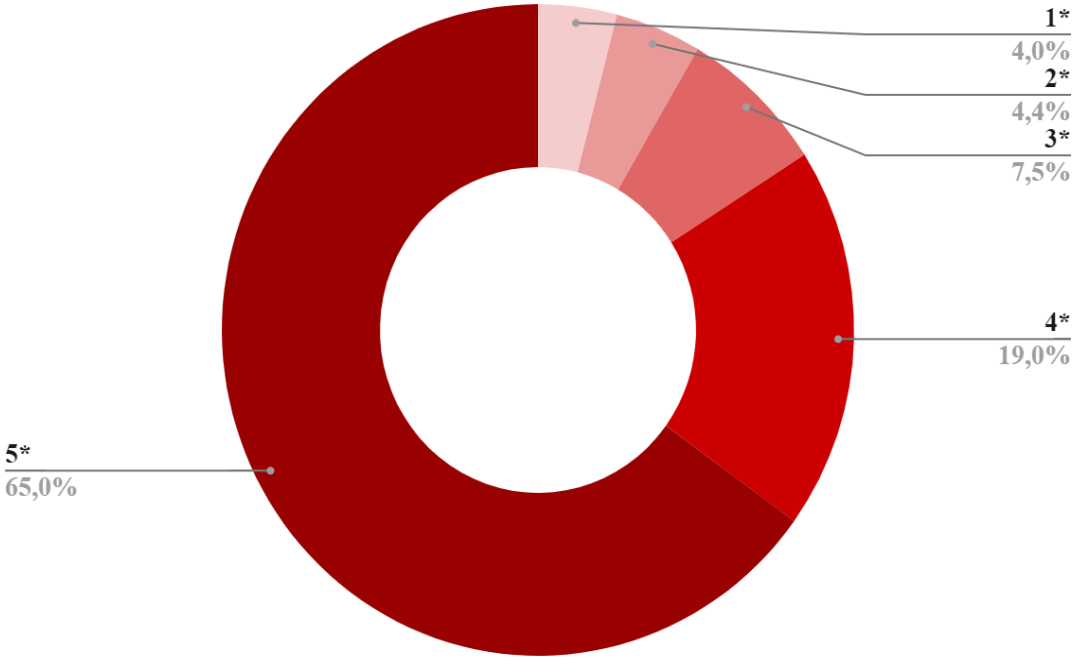

Figure 13: Review allocation according to the rating in the Train dataset (Source: Authors)

The data set's sentiment spectrum is a 0–2 rating scale, and Figure 14 shows how the data set's distribution of comments by sentiment is distributed for each category of sentiment. with the majority of user evaluations being favorable (sentiment = 2). containing 55930 comments.

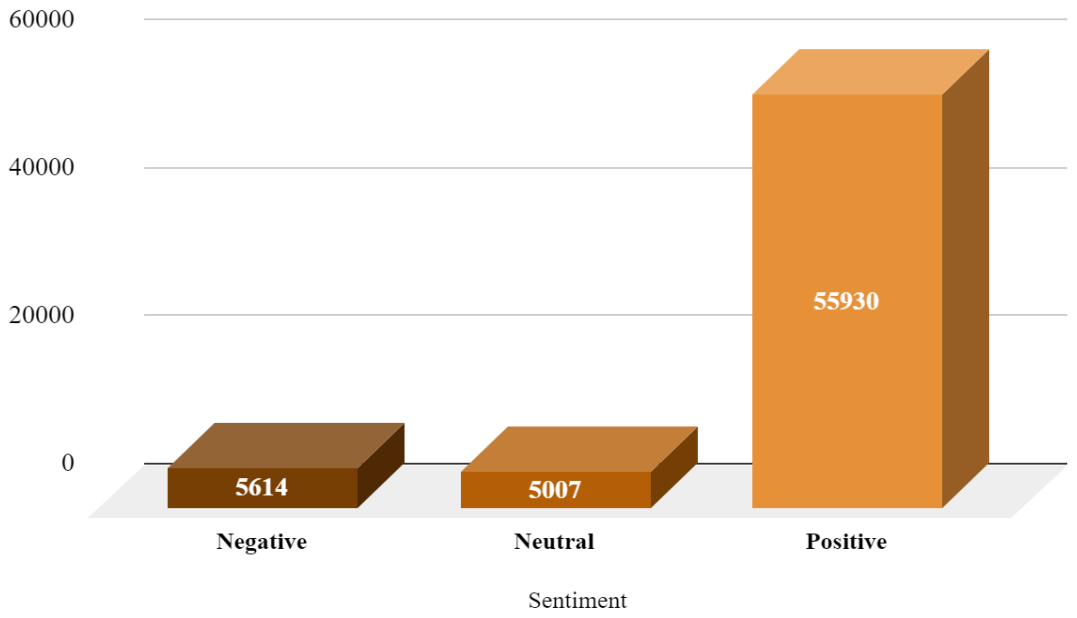

Figure 14: Review distribution according to polarity in the Train dataset (Source: Authors)

The boxplot chart in Figure 15 shows the distribution of word_count by each sentiment point. Word count allocation in a comment based on sentiment doesn't make too big of a difference:

- For a polarity 0 (negative), the number of words per comment measured in units of measure recorded values from about 5 to 390. The box indicates a value range from about 50 to 200, with median is at the left skewed position. There are some outliers marked as single points scattered from about 300 to 3300. All of these values are outliers and are not within the range allocated in the box.
- For a polarity 1 (neutral), the number of words per comment measured in units of measure records values from about 7 to 350. The box indicates a value range from about 40 to 190, with values The median is at the left skewed position. There are some outliers marked as single points scattered from about 320 to 3700. All these values are outliers and are not within the range of values allocated in the box.
- For a polarity 2 (positive), the number of words per comment measured in units of measure recorded values from about 5 to 300. The box indicates a value range from about 40 to 120, with a value The median is at the left skewed position. There are some outliers marked as single points scattered from about 250 to 2400. All these values are outliers and are not within the range of values allocated in the box.

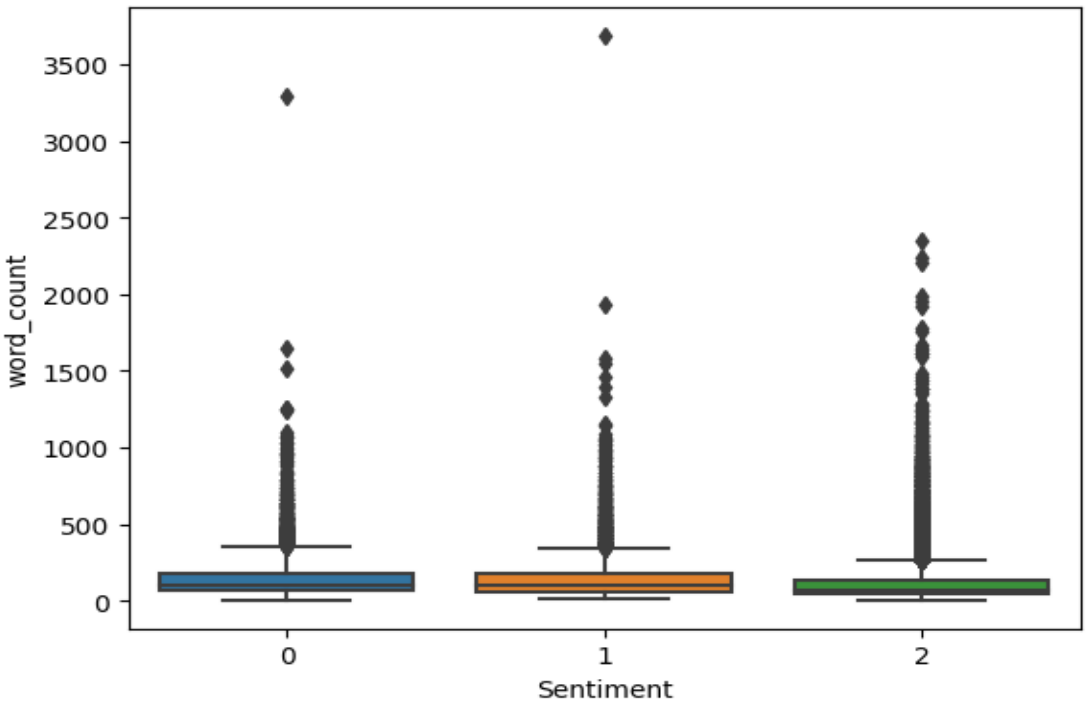

Figure 15: Allocate word count of each review according to polarity (Train dataset) (Source: Authors)

The train dataset's word cloud in Figure focuses on words like "resort," "hotel," "restaurant," “room”. This illustrates how users are frequently drawn to these types of lodgings.

Figure 16: Word cloud of the Train dataset (Source: Authors)

Words frequently used in user comments are categorized by each sentiment level of the data collection and are displayed in Figure 17, 18 and 19. When concentrating on words like hotel, stayed, location, and so on, word clouds do not show significant differences in sentiment levels.

Figure 17: Word cloud of negative review

Figure 18: Word cloud of neutral review

Figure 19: Word cloud of positive review

(Source: Authors)

In general, the data set is imbalanced with the number of positive reviews nearly 10 times higher than the negative or the neutral ones. Therefore, data sampling methods should be taken into consideration, which will be described in the next section.

3.2.3. Modeling

In this step, the recurrent neural network-based LSTM and pretrained model distilBert are taken into account for classification task. About LSTM, two training dataset, including original dataset and under-sampled dataset, are used in order to evaluate the impact of imbalanced training dataset.

Three models, including LSTM with original dataset, LSTM with under-sampled dataset and distilBert are developed. These model are considered Model 1, Model 2 and Model 3 , respectively.

3.2.3.1. Model 1

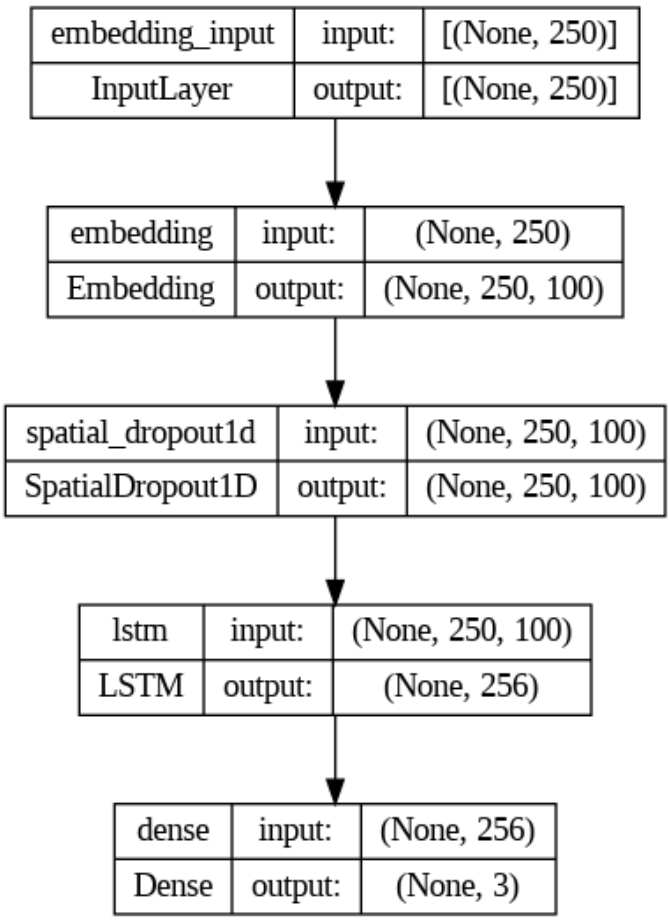

Figure 20: The Architecture of Model 1 and Model 2 (Source: Authors)

The architecture of Model 1 and Model 2 consists of 4 layers as illustrated in Figure 20. Each layer is described in more detailed as below.

The Embeding layer is established with the size of vocabulary at 50000, embedding dim at 100 and max length of a sequence (max number of words in each review) at 250.

The SpatialDropout1D is applied to reduce overfitting by ignoring a number of neurons which is set at 0.3 in this study. To be more specific, 30% of the total neurons receives zero no matter what activation function is used.

Then LSTM and Dense layers are deployed with Adam Optimizer, loss as categorical cross entropy. Adam Optimizer is considered as on of ther most successful optimizer algorithm by estimating the adaptive learning rate for all parameters involved in the training of gradients. Thus, it can reduce computational cost, memory, and time needed to complete training process. In addition, accuracy is

chosen to observe training performance of models and assess models' output. The model is compiled with 2 epochs and batch size at 64.

3.2.3.2. Model 2

Model 2 applies exactly the same architecture of the Model 1 with the difference in data train. As mention in the EDA step, data train is extremely imbalance with the number of positive review accounts for over 84%, which is high likely to cause bias phenomenon. Therefore, resample dataset should be taken into account.

Because of the efficient data and to reduce time of training, under-sampling method is selected. As a result, the data train used to train this model has the equal number of review from each polarity with the total at 16809 instances.

3.2.3.3. Model 3

The third model leverages pre-trained model distilBert, a small, fast, cheap and light Transformer model based on the BERT architecture. DistilBert consumes significantly more time and huge memory, compared to the last two models. Due to the GPU memory limit and a shortage of time as well as to avoid bias phenomenon, the data train is under-resampled and the final size of data train is 16,638. Also, the limit of GPU session time allows only 1 epoch to be complied and the batch size to be determined by 12.

4. Result

This section provides a discussion of the experimental results. In this study, three algorithms are evaluated in terms of accuracy, precision, recall and F1 score for each polarity. Weighted average scores are applied to compare three models because they take into account the porlarity's ratio, which is significant in such an impalanced data test.

4.1. Model 1

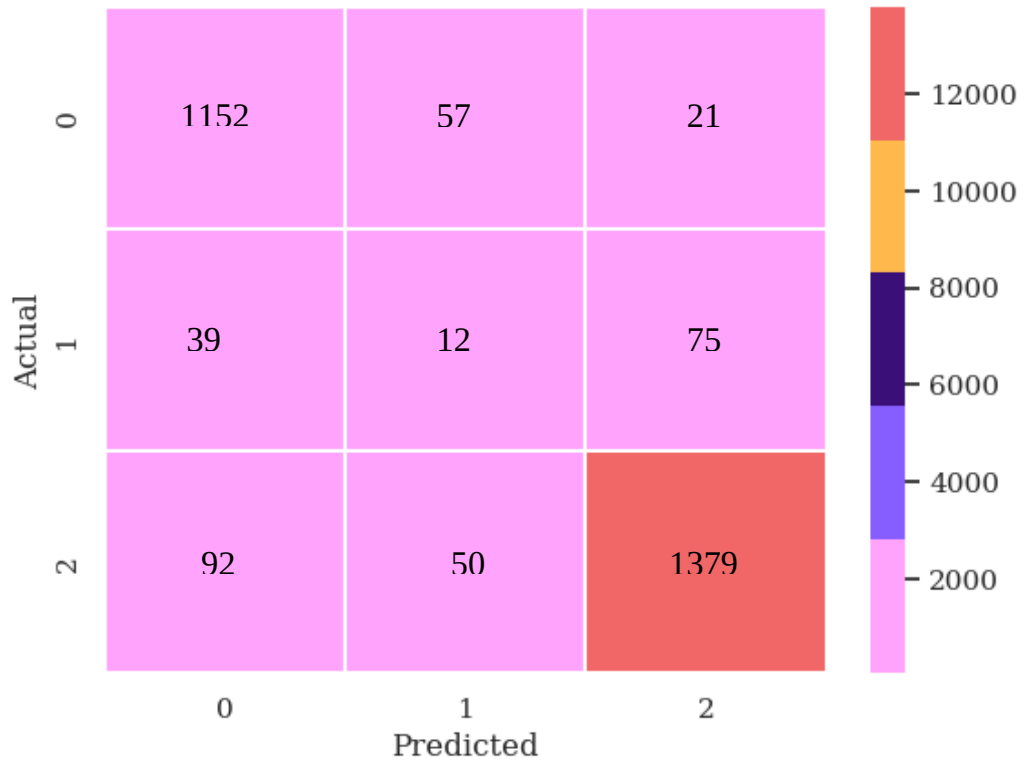

Figure 21: Confusion matrix of LSTM (Source: Authors)

Table 3: Classification report of Model 1 (Source: Authors)

	Precision	Recall	F1-Score	Support
0	0.70	0.81	0.75	1428
1	0.53	0.10	0.16	1269
2	0.93	0.99	0.96	13941

accuracy			0.91	16638
macro avg	0.72	0.63	0.63	16638
weighted avg	0.88	0.91	0.88	16638

With the polarity 0, this model shows that all measures of Precision, Recall and F1-Score are 70%, 81% and 75%, respectively. The result of polarity 2 also indicates that all Precision, Recall and F1-Score measures on negative reviews is 93%, 99%, 96%, the percentage of the indications tending towards negative reviews is much larger than the other two polarities, this demonstrates that model 1 seems biased. This bias phenomenon can be explained by the imbalance data train with positive reviews making up over 84%.

What's more is the far lower score when analysing neutral reviews, compared to the others. The poor performance in both precision and recall shows the model's limit in the ability to determine neutral reviews.

4.2. Model 2

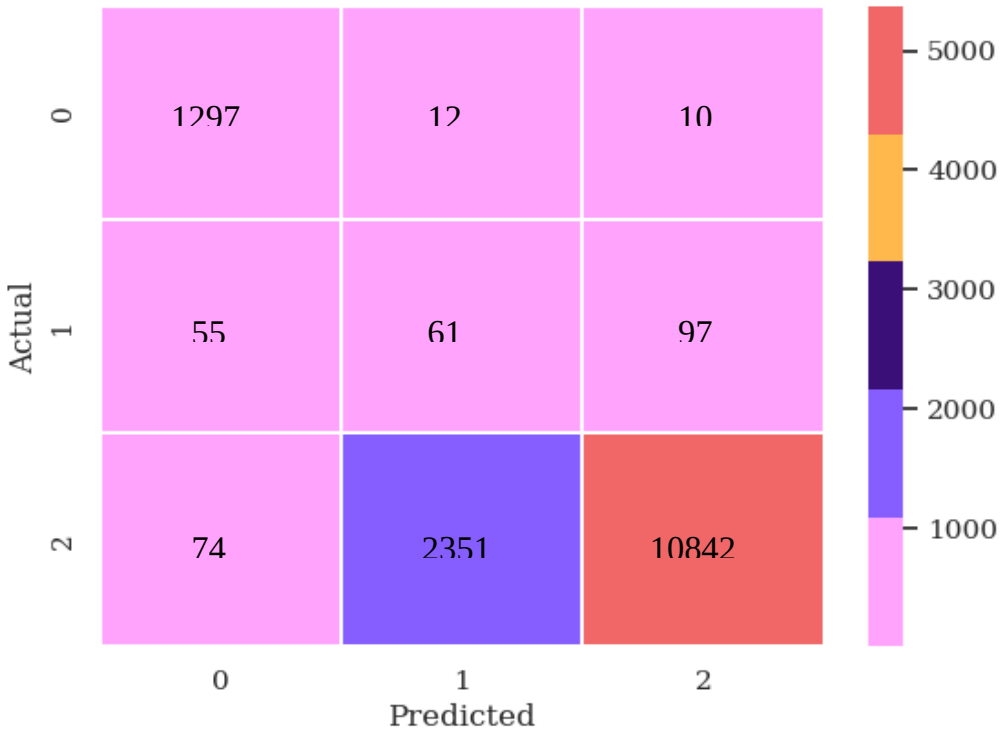

Figure 22: Confusion matrix of Model 2 (Source: Authors)

Although it gives lower accuracy, under-sampling helps reduce bias in the positive polarity. In the under-sampling model, the performance of the neutral class is improved better than that of Model 1. But this model raises another problem in confusing positive reviews with the others, leading to low recall score.

Table 4: Classification report of Model 2 (Source: Authors)

	Precision	**Recall**	**F1-Score**	**Support**
0	0.50	0.91	0.64	1428
1	0.20	0.49	0.28	1269
2	0.99	0.78	0.87	13941
accuracy			0.77	16638
macro avg	0.56	0.72	0.60	16638
weighted avg	0.89	0.77	0.81	16638

With the polarity 0, this model shows that measures of Precision is 50%, Recall is 91% and F1-Score is 64%. However, with the polarity 1, the measures of Precision, Recall and F1-Score are more lower with the results are 20%, 49% and 28%. This has changed rapidly in the case of polarity 2, the ratio with 99%, 78% and 87% belong to the measures of Precision, Recall and F1-Score respectively. Bias is more limited in this model than in model 1, which helps to lessen the gap of Precision, Recall, and F1-Score between polarity.

4.3. Model 3

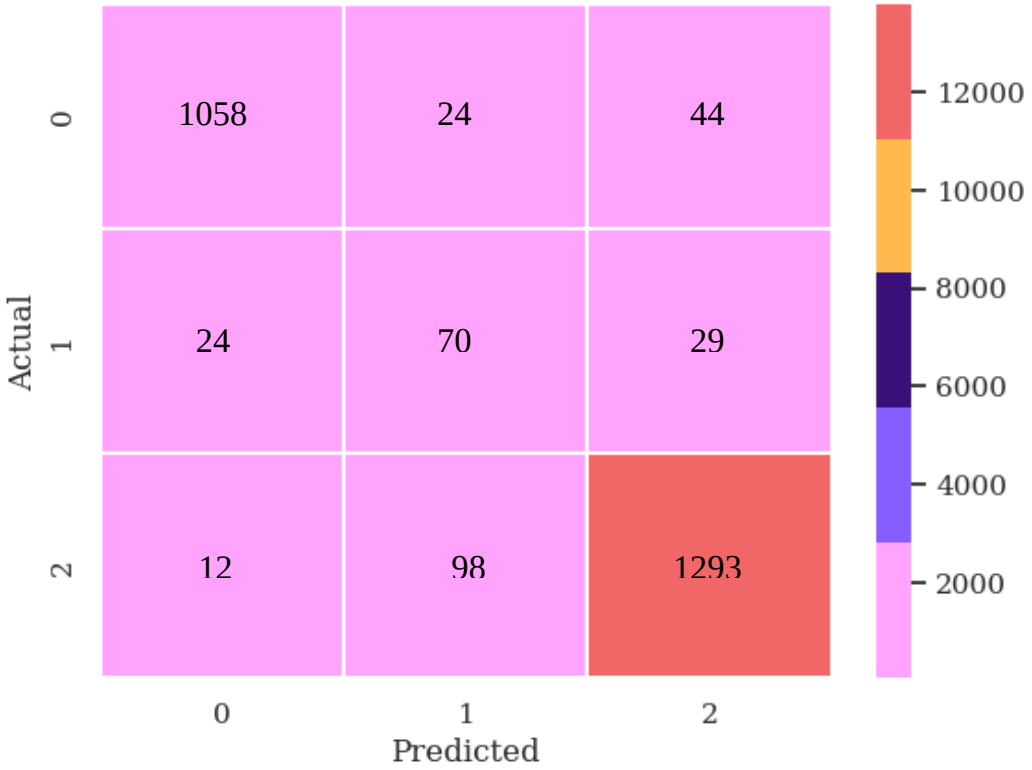

Figure 23: Confusion matrix of Model 3 (Source: Authors)

The model achieved an accuracy of 88%, indicating that it correctly classified 88% of the samples in the dataset. This performance measure is a good indication of the reliability of the model's predictions, however, it is also important to check other metrics like Precision, Recall and F1-score.

For polarity 0 (negative), this model shows that all measures of Precision, Recall and F1-Score are 74%, 79% and 76%. The measures of Precision, Recall and F1-Score have the highest results with the polarity 2, at 97%, 92% and 95% respectively. In contrast, the ability to predict negative polarity experienced the lowest performance in all measure metrics mentioned above.

Table 5: Classification report of Model 3 (Source: Authors)

	Precision	**Recall**	**F1-Score**	**Support**
0	0.74	0.79	0.76	1343
1	0.36	0.57	0.44	1245

2	0.97	0.92	0.95	14050
accuracy			0.88	16638
macro avg	0.69	0.76	0.72	16638
weighted avg	0.91	0.88	0.89	16638

4.4. Comparison

Table 6: Performance metrics of 3 models (Source: Authors)

Model	Precision	Recall	F1-Score	Accuracy
Model 1	0.88	0.91	0.88	0.91
Model 2	0.89	0.77	0.81	0.77
Model 3	0.91	0.88	0.89	0.88

Table 6 compares the performances of Model 1, Model 2, and Model 3 classification techniques with all measures of Precison, Recall, F1-Score, Accuracy. The effect of each polarity are measured by Precison, Recall and F1-Score metrics and the overall classification effect of the model was measured by the Accuracy.

Compared to Model 2 and Model 3, the overall accuracy of the Model 1 was 91%, which is higher. In terms of classifications, the Model 1 provides the best outcomes across all three categories. According to the table, the Model 1 findings appear to be higher due to its tendency for bias problems with imbalanced data. Since Model 3 accuracy of 88% is significantly higher than that of Model 2 and it also seems to be reduced bias issues that the Model 1 has, it achieves good results in all three models. This results in the Model 3 most pricise prediction. Model 3 delivers the best results in this dataset, but this does not indicate that it is generally superior or inferior.

It just contributes to demonstrate that the Model 3 is better suited for the dataset chosen for the current research.

All the models share a limilarity in far poorer performance for classify neutral reviews, compared to the others. This implicates the limited ability to detect neutral reviews of all three models. Therefore, there is a need to focus more on this polarity, including labeling more carefully instead of just basing on rating star.

To examinine the classification result for several specific reviews, a Flask App is deployed. Sample reviews are inputted and then their polarities analysed by three models are shown as illustrated in Figure 24. In some cases, the same review receives different results from 3 models, especially in neutral reviews that contain both negative and positive words.

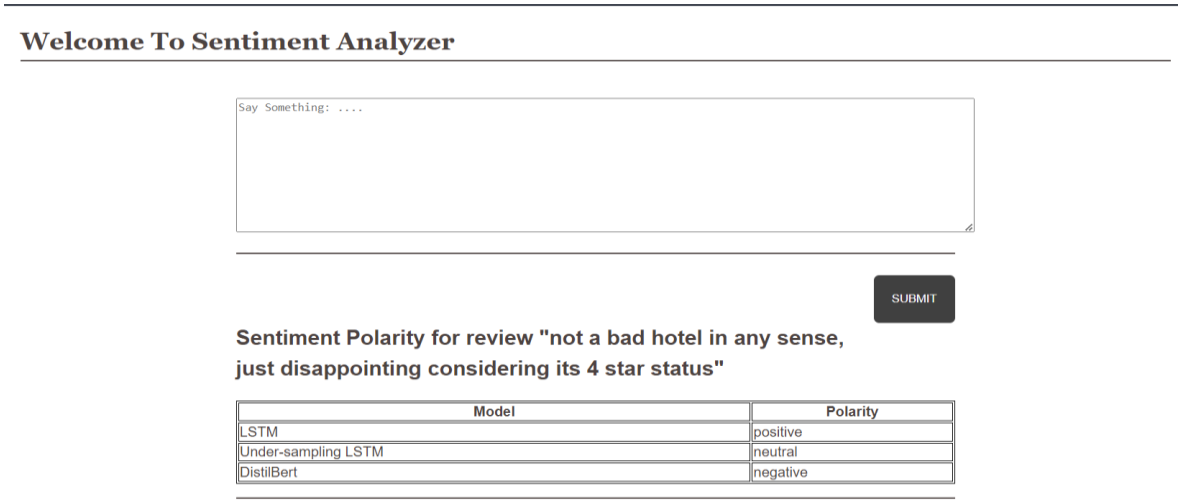

Figure 24: Flask App to show results of classification task (Source: Authors)

5. Conclusion and further research

Analysing online reviews has become an important task to discovery customer satisfaction and assess the operational performance of hotels. This study proposed three models, including LSTM trained with original data, another one with uder-sampled data, and distilBert model. Among these models, distilBert achieves the highest F-score, at 0.89, while the highest accuracy sore at 0.91 belongs to the LSTM model trained with original dataset. Due to highly imbalanced data, F1-score is taken more seriously than the accuracy score. The better performance of distilBert model is also observated via experiments on Flask App. The study, to some extent, has achieved its objective, classifying customer reviews into positive, neutral or negative. In addition, from the review dataset of all hotel in Vietnam, exploratory analysis shed a light on some factors contributing to negative experience (like staff, check-in, etc) and some help hotels meet customer satisfaction (beach, location, etc). This motivates to further research carried on aspect level to give more clear and precise assessment of each aspect of each hotel and a overview of hotel's operational status in Vietnam, contributing managerial implications.

However, there are also some lacks of this study we can see. Firstly, translating data into English loses the original meaning of it, causing the problem that the meaning may be changed. To solve this problem, the model should be trained in the original language to keep the meaning exactly the same. Furthermore, this study analyzes the sentiment based on the entire individual review of the dataset. To expand, the researchers can level up the model to dissect the reviews in order to promote the proposes from the result for business application. Another restriction is the limited time and technology. Because the study is done in a short period of time and the shortage of GPU used to train the model, the dataset is not made full use of. To fix this shortage, the researchers can optimize the algorithm and upgrade the technical basis for higher performance.

REFERENCES

A. Gandomi, and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, Vol.35, No.2, pp.137-144, 2015.

A. Ghazian, M. Hossaini, and H. Farsijani, “The effect of customer relationship management and its significant relationship by customers’ reactions in LG Company”, In: Proc. of International Conf. on Economics and Finance, pp.42 – 50, 2016.

Agarwal, S. (2022). Deep learning-based sentiment analysis: Establishing customer dimension as the lifeblood of business management. Global Business Review, 23(1), 119-136.

Anis, S., Saad, S., Aref, M. (2021). Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques. In: Hassanien, A.E., Slowik, A., Snášel, V., El-Deeb, H., Tolba, F.M. (eds) Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020. AISI 2020. Advances in Intelligent Systems and Computing, vol 1261.

Arora, A. M. H., & Bansal, B. D. M. (2022). LSTM and Bi-LSTM Deep Learning Technique for better Tourism Services in future by analyzing Hotel Reviews. JOURNAL OF ALGEBRAIC STATISTICS, 13(3), 3114-3123.

Arora, H., & Bansal, M. (2022, November). Comparative analysis of hotel reviews using proposed LSTM based deep learning model with classical machine learning techniques. In AIP Conference Proceedings (Vol. 2481, No. 1, p. 020003). AIP Publishing LLC.

Bhoi, D., Thakkar, A., & Patel, R. (2023, March). Sentiment Analysis Using an Improved LSTM Deep Learning Model. In Third Congress on Intelligent Systems: Proceedings of CIS 2022, Volume 1 (pp. 249-262). Singapore: Springer Nature Singapore.

E. Park, J. Kang, D. Choi, and J. Han, “Understanding customers' hotel revisiting behaviour: a sentiment analysis of online feedback reviews”, Current Issues in Tourism, pp.1-7, 2018.

H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review", Expert Systems With Applications, Vol.118, No.3, pp.272-299, 2019.

Hossen, M. S., Jony, A. H., Tabassum, T., Islam, M. T., Rahman, M. M., & Khatun, T. (2021, March). Hotel review analysis for the prediction of business using deep

learning approach. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) (pp. 1489-1494). IEEE.

Ilham Esa Tiffani, “Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review”, J. Soft Comput. Explor., vol. 1, no. 1, pp. 1-7, Oct. 2020.

Intellipaat, 2023. “What is LSTM? Introduction to Long Short Term Memory”. Retrieved on March 28th, 2023 from: https://intellipaat.com/blog/what-is-lstm/

Khotimah, D. A. K., & Sarno, R. (2019). Sentiment analysis of hotel aspect using probabilistic latent semantic analysis, word embedding and LSTM. International Journal of Intelligent Engineering and Systems, 12(4), 275-290.

M. Ady and D. Quadri-Felitti, "Consumer Research Identifies Which Attributes Are Most Important to Travelers When Booking A Hotel", Trust You, London, 2015

M. Geetha, P. Singha, and S. Sinha, "Relationship between customer sentiment and online customer ratings for hotels - An empirical analysis", Tourism Management, Vol.61, No.22, pp.43-54, 2017.

Maw, S. Y., & Khine, M. A. (2019). Aspect based Sentiment Analysis for travel and tourism in Myanmar Language using LSTM (Doctoral dissertation, MERAL Portal).

Muhammad, P. F., Kusumaningrum, R., & Wibowo, A. (2021). Sentiment analysis using Word2vec and long short-term memory (LSTM) for Indonesian hotel reviews. Procedia Computer Science, 179, 728-735.

N. Akhtar, N. Zubair, A. Kumar, and T. Ahmad, "Aspect based Sentiment Oriented Summarization of Hotel Reviews", In: Proc. Computer Science, Vol.115, pp.563-571, 2017.

Öhman, E. (2021, December). The Validity of Lexicon-based Sentiment Analysis in Interdisciplinary Research. In Proceedings of the Workshop on Natural Language Processing for Digital Humanities (pp. 7-12).

Ray, B., Garain, A., & Sarkar, R. (2021). An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. Applied Soft Computing, 98, 106935.

Shi, H. X., & Li, X. J. (2011, July). A sentiment analysis model for hotel reviews based on supervised learning. In 2011 International Conference on Machine Learning and Cybernetics (Vol. 3, pp. 950-954). IEEE.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

Y. Hu and K. Chen, “Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings”, International Journal of Information management, Vol.36, No.3, pp.929-944, 2016.

APPENDIX

Appendix A. Teamwork Measurement

Student ID	Name	Role	Responsibility	Contribute
K204110566	Tran Nguyen Ngoc Hieu	Member	- Data Exploratory Analysis - Write Chapter 3 (EDA) - Write Chapter 4 (Result of 3 models) - Design slides	20%
K204110605	Nguyen Thanh Phat	Member	- Text cleaning - Write Chapter 1 - Write Chapter 5 - Format word - Design slides	20%
K204110611	Bang Nhan Tri	Member	- Text cleaning - Write Chapter 2 - Write Chapter 3 (Research process, Text Cleaning) - Design slides	20%
K204111797	Nguyen Thi Linh Anh	Member	- Write Chapter 2 - Train model 1&2 - Write Chapter 4 (Comparison) - Design slides	20%
K204111813	Tran Thi Kieu Trinh	Leader	- Train model 3 - Run Flask App	20%

			- Write Chapter 3 (Modeling) - Write Chapter 4	