CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

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

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

5.2 Future Works

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

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

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