Master Thesis

(MORS) : Multi-Output Regression System for Analyzing Airline Customer Satisfaction: A Skytrax Dataset Study

Theodoros Giannilias

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science of Artificial Intelligence
at the Department of Advanced Computing Sciences
of the Maastricht University

Thesis Committee:

Professor Jerry Spanakis

Maastricht University
Faculty of Science and Engineering
Department of Advanced Computing Sciences

August 22, 2023

Contents

List of Abbreviations

Seat Comfort: SC

Cabin Staff Service: CSS

Food & Beverages: FB

Ground Service: GS

Inflight Entertainment: IFE

Value For Money: VFM

Wi-fi & Connectivity: Wi-Fi

User Rating: UR

Abstract

With the advent and increased use of the internet, social media has become an integral part of people's daily routine, eventually leading to social platforms being huge data sources. Nowadays, the airline industry operates in a highly competitive market, and customer satisfaction plays a crucial role in their success. In that sense, it is significant to understand the context that passengers evaluate airline services and to recognize their most valued dimensions of satisfaction. This paper aims to develop a Multi Output Regression System (MORS) to assist the aviation industry in identifying patterns and correlations between several aspect scores, based on the customer reviews of the Skytrax dataset. The paper implements some baseline model architectures, which predict every score based on the customer reviews of the Skytrax dataset, and tries to combine those to demonstrate their influence in MORS. Additionally, the research focuses on showing the key factors that determine the customer's ratings and the evaluation challenges of this process. It is just as important to mention, that this work emphasizes the importance of incomplete reviews and the way they affect the user's overall rating score.

Acknowledgements

I would like to thank Professor Jerry Spanakis for the valuable scientific guidance he has provided me throughout the course of this thesis, as well as for the immediate help he offered me whenever it was needed. In addition, I would like to thank the professors of the Department of Advanced Computing Sciences, who have given us the knowledge we need in order to become ourselves, the ones who eventually will try to evolve this scientific field. Afterwards, I would like to thank Maastricht University, the Department of Advanced Computing Sciences and more specifically the Artificial Intelligence M.Sc. program for providing me the necessary equipment to accomplish this thesis. Finally, I would like to thank my family and my friends for their support and encouragement throughout all these years.

Chapter 1

Introduction

People's communication styles have greatly changed as a result of the rapid progress of new media and information technology, moving more toward sharing thoughts and expressing emotions on social media and open forums (Song, Guo, & Zhuang, 2020). Social media has become an essential component of people's everyday lives as a result of the development and widespread usage of the internet, which has given rise to enormous user-generated content data sources. Numerous industries, including aviation, now have unprecedented potential to access the abundance of data available on social media platforms and acquire important insights into consumer experiences and preferences as a result of the data revolution.

In a highly competitive market, consumer loyalty and happiness have become crucial in the aviation sector. By providing great services and individualized attention to each passenger, airlines work to stand out from their competitors (Sezgen, Mason, & Mayer, 2019). However, because of the enormous amount of data and the complicated nature of consumer feedback, comprehending how passengers perceive and assess airline services has remained a challenging task (Park, Lee, & Nicolau, 2020). Airlines now have new ways to understand client attitudes and sentiments concealed in textual data thanks to the development of sentiment analysis and natural language processing (NLP) tools.

Given these possibilities, the primary objective of this thesis is to create a sophisticated **Multi-Output Regression System (MORS)** utilizing a modified Skytrax dataset (`https://surfdrive.surf.nl/files/index.php/s/k0QsJYib6Moh4e2`). The MORS attempts to enhance forecasting of the individual and the overall user score of passengers' travel experiences with aviation businesses. This unique approach aims to outperform the performance of distinct baseline models by using the MORS, which incorporates all pertinent scores into the learning process.

Additionally, this work addresses a variety of diverse and multidimensional research concerns. The study investigates the main elements that influence a customer's evaluation of an airline's aspect and considers the difficulties in assessing the grading procedure based on these elements.

To do this, a thorough study of coefficients and feature relevance in both the baseline models and the MORS is carried out to identify the important key variables that significantly affect consumers' rating decisions. The thesis also establishes the relationship between these scores and examines the impact of integrating all pertinent scores for various characteristics on the user's overall evaluation of an airline. Aspect ratings are frequently treated as independent entities in traditional methods, which ignores their potential interdependencies. The MORS, in

contrast, considers these relationships, revealing how each component affects the overall customer rating as well as how they relate to one another.

1.1 Research Questions

This Master thesis addresses the following primary research questions :

1. What are the key factors that determine a customer's rating of an airline's aspect, and what challenges are associated with evaluating the grading process based on these factors?
2. What is the impact of combining all relevant scores for different aspects on each score including the user's overall rating of an airline, and how do these scores relate to each other?
3. How can incomplete reviews with missing data for some aviation aspects be handled, and how does such missing data affect the prediction of the user's overall rating?

1.2 Main Thesis Contribution

The main contributions of this master thesis research project are:

1. **Multi Output Regression System (MORS) for Enhanced Performance**: This study presents a sophisticated (MORS) that incorporates all relevant scores into the learning process, using the defined baseline models as a Multi Output System. By combining data from various pertinent scores, this novel approach aims to determine whether the performance of baseline models can be enhanced.
2. **Identification of Significant Key Factors for Customer Ratings**: This research employs rigorous coefficient and feature importance analyses in both the baseline and multi-output regression models. By examining the magnitude of the weights and extracting top features, this thesis identifies significant key factors that strongly influence customers' decisions.

1.3 Upcoming Chapters

The structure of this thesis includes the following chapters:

- **Chapter II - Related Work:** This section will provide a comprehensive overview of the existing literature and related works in the field of online review analysis with statistical models, neural network-based approaches, and state-of-the-art solutions for aviation industry.
- **Chapter III - Proposed Methodology:** This section explores word embeddings that were used and their application in converting textual data into numerical representations suitable for machine learning models. It also examines the baseline model architectures used in the project, highlighting their strengths and limitations regarding both the rating prediction and the coefficient/feature importance analysis tasks. Finally, the main contribution of this work Multi-Output Regression System (MORS) is described by analyzing the sophisticated approach that incorporates all relevant scores into the learning process, providing prediction enhancement and valuable insights into customer decision-making.

- **Chapter V - Data:** This section will describe the modified dataset, the insufficient issues & challenges that occurred as well as the pre-processing steps in detail, ensuring transparency and reproducibility of the experiments.
- **Chapter VI - Experimental Evaluation:** In this section, we will detail the experimental setup, elucidate the conducted experiments along with their outcomes, and highlight the performance disparities between the baseline models and the **MORS**. Additionally, this work offers a comprehensive presentation of the pivotal indicators that substantially shape customers' opinions. This will be accomplished through a rigorous coefficient/feature importance analysis. The pivotal indicators' influence will be underscored in the subsequent analysis section.
- **Chapter VII - Conclusion:** The final section will state the general conclusion of this work, discuss the implications of the research findings and outline potential future work in this domain.

Chapter 2

Related Work

Several noteworthy studies have been carried out in the area of comprehending client feedback in similar contexts of this Master Thesis. The field of sentiment analysis and opinion mining has evolved significantly in recent years, moving from traditional two- or three-class sentiment analysis to more fine-grained approaches that aim to capture the complex inner world of human emotions.

2.1 Online Review Analysis with Statistical Models

Prior research has focused on the significance of review helpfulness and reviewers' expertise in influencing users' perceptions of online ratings in the setting of online reviews. In the future, we can identify users' happiness with various service attributes using attribute extraction techniques.

(Liu & Park, 2015) studied the impact of review features and characteristics of providers on the perceived utility of online consumer reviews. They employed the TOBIT regression model to analyze the data, which was suitable for handling non-negative and skewed dependent variables like useful votes, including cases with zero votes.

(Fang, Ye, Kucukusta, & Law, 2016) conducted a two-level empirical investigation, analyzing factors influencing review value at both review and reviewer levels. They used the negative binomial model to assess helpfulness votes, capturing greater variance, and the Tobit II model to estimate average helpfulness, accounting for both censored and uncensored data.

Furthermore, social media platforms and microblogs are vital data sources in sentiment analysis. (Pak & Paroubek, 2010) conducted sentiment analysis and opinion mining on Twitter, creating a corpus categorized by specific emoticons. They employed linguistic analysis to identify sentiment patterns in POS-tag distributions. Using multinomial Naive Bayes, they built a sentiment classifier using POS tags and N-grams as features for efficient sentiment determination in microblogging data. In affective computing, (Li, Lin, Lin, Wang, & Meng, 2018) provides a comprehensive survey of text-based emotion analysis. It explores lexicon-based and machine learning-based computational approaches and investigates emotional models, including categories and dimensions. The machine learning approach utilizes linguistic features for training emotion classification algorithms, improving system accuracy, while the lexicon-based approach relies on emotion lexicons with annotated emotional words.

(He, Lee, Ng, & Dahlmeier, 2019) proposed IMN, an Interactive Multi-task Learning Network for aspect-based sentiment analysis (ABSA), combining aspect-level sentiment classification and aspect and opinion term co-extraction. It features a novel message-passing system for informative interactions between tasks, transmitting information and correlations across components. Leveraging diverse data sources and incorporating fine-grained token-level tasks with document-level labeled corpora, IMN captures domain-specific information, enhancing ABSA performance through multitask learning and interactions.

(Eslami, Ghasemaghaei, & Hassanein, 2018) proposed a theoretical model to study factors impacting perceived review helpfulness. They analyzed user-generated reviews, considering review length, score, and argument frame as potential factors. The model was validated using sentiment analysis, PLS-SEM, ANOVA, and an ANN to predict review helpfulness.

2.2 Neural Network Based Approaches

When it comes to deep networks, (Tang, Qin, & Liu, 2016) aimed to enhance aspect-level sentiment classification using a deep memory network with neural attention models to capture context word importance. Overcoming LSTM-based model limitations, the approach achieved results comparable to feature-based SVM systems on SemEval 2014 datasets, significantly advancing aspect-level sentiment analysis.

(Tang, Qin, Feng, & Liu, 2016)developed target-specific long short-term memory (LSTM) models for target-dependent sentiment classification, incorporating target information to capture the semantic relatedness between target and context words. The extension to TD-LSTM and TC-LSTM allowed a better understanding of target-context relationships, achieving state-of-the-art performance without external lexicons or parsers.

(Chen, Zhuo, & Ren, 2019) proposed a modified Gated Recurrent Unit for sentiment classification(GRNN-SR), utilizing a multiplicative technique to model sentimental relations with two hidden states for sentiment polarity and sentiment modifier context. The (GRNNSR) (Santur, 2019) (Sachin, Tripathi, Mahajan, Aggarwal, & Nagrath, 2020) showcased its potential as a promising approach for sentiment analysis tasks without the need for language characteristics or intricate linguistic regularization.

An attention-based LSTM model for aspect-level sentiment categorization is presented by (Wang, Huang, Zhu, & Zhao, 2016). The model learns to differentiate between the weights of various words in a phrase by including attention mechanisms, concentrating on the most crucial elements for sentiment analysis. The performance and interpretability of aspect-level sentiment classification are improved by this method.

(Dereli & Saraclar, 2019) utilized annual reports to predict the stock return volatility of publicly traded companies. They adopted a convolutional neural network (CNN) model with word embeddings instead of a manual financial sentiment lexicon, effectively capturing context from the reports. The multichannel embedding layer and transferred convolution layer enabled better performance than lexicon-based models, highlighting the potential of word embeddings in financial text regression research for more accurate volatility predictions. (Giannakopoulos, Musat, Hossmann, & Baeriswyl, 2017) introduced a novel Aspect Term Extraction (ATE) method, combining supervised and unsupervised learning due to limited datasets. They utilized a B-LSTM & CRF classifier for token-based classification and feature extraction, achieving accurate aspect term identification in raw opinion texts. Their unsupervised approach outperformed the supervised baseline, advancing ATE methods.

Several papers focused on evaluating the performance of fancier models, including transformers (Joshy & Sundar, 2022) in sentiment analysis tasks.(Talaat, 2023) developed a framework using two BERT models (RoBERTa and DistilBERT) combined with BiGRU and BiLSTM layers to create eight hybrid models for predicting emotions in tweets. The proposed models achieved higher accuracy in sentiment classification than BERT alone and outperformed classical machine learning models.

2.3 State of the art solutions on aviation industry

This paper (Charatsaris, 2017), which this thesis draws inspiration from, builds upon prior aviation industry research from the Skytrax dataset and introduces a comprehensive approach to Aspect Based Sentiment Analysis (ABSA). The approach involves target identification, category detection, and sentiment analysis. Effective techniques like BiLSTM+CRF and CNN+BiLSTM+CRF are employed for precise target identification. SVM and BERT are selected as potent text classification models for category detection. BERT is employed in sentiment analysis to predict sentiment polarity, considering both target and aspect categories, leading to comprehensive insights into customer sentiments in airline reviews.

(Tiwari et al., 2018) conducted sentiment analysis of aviation services using a Twitter dataset, association rules to mine frequent item sets, and a space vector model to translate text into space vector and then into BIRCH clustering. According to the data, the appearance of "flight" in negative comments is generally accompanied by the appearance of "cancel," "delay," and "late," indicating that users' thoughts are negative.

Additionally, (Sezgen et al., 2019) applied Latent Semantic Analysis (LSA) to 5120 airline passenger reviews from TripAdvisor.com. LSA is a mathematical text-mining technique that uncovers patterns in unstructured data. The analysis identified factors influencing passenger satisfaction and dissatisfaction in various airline business models and service classes.

(Park et al., 2020) analyzed 157,035 TripAdvisor reviews to study airline service quality's impact on satisfaction. They used the Tobit model for constrained data and identified asymmetrical effects of service attributes on satisfaction, validating Herzberg's theory. Practical implications were provided for airline executives during Covid-19.

Furthermore, (Song et al., 2020) analyzed user reviews from the Skytrax dataset, particularly on flight delays, using text mining and lexicon-based sentiment analysis. Their goal was to understand passengers' emotional responses and concerns during flight delays by extracting valuable patterns from unstructured text data and obtaining sentiment ratings. This approach provided intuition about customer satisfaction and how much flight delays affected passengers' sentiments.

(Verma & Davis, 2021) introduced a two-level technique for extracting implicit aspects and opinions from airline reviews. They used conditional random fields (CRF) with stochastic gradient descent and L2 regularization for entity classification and ensemble learning methods like Voting Classifier and XGBOOST for implicit aspect classification. Overcoming class imbalance, they achieved superior performance compared to previous valuable approaches.

The examination of sentiment analysis, aspect extraction, customer satisfaction analysis, and text review regression similarity of this work is aided by the combined efforts of these studies. Based on the above-stated research that has been done, the path of applying Multi-Output Regression to combine all the relevant scores of the Skytrax dataset has not been pursued, which is why this paper aims to follow this approach as a research subject.

Chapter 3

Proposed Methodology

The "Proposed Methodology" portion of this thesis offers a thorough strategy for precisely forecasting various scores using the text evaluations from the Skytrax dataset. To do this, it blends machine learning, deep learning, and transformer model architectures.

To improve the performance of machine learning models like **Ridge Regression** and **Support Vector Regression(SVR)**, the process begins with **Term Frequency - Inverse Document Frequency (TF-IDF)** word embeddings. Additionally, deep learning models like **Convolutional Neural Network(CNN)**, **Gated Recurrent Unit (GRU)**, and **Lonst-Short Term Memory (LSTM)** networks are combined with pre-trained **Global Vectors for Word Representation (GloVe)**, **FastText**, and **Custom Word2Vec** word embeddings to provide contextualized representations of the reviews.

Distilation BERT (DistilBERT), a potent transformer-based model with an associated pretrained tokenizer, is included to support complex language processing jobs. To determine the factors that most significantly influence consumer opinions, each model is subjected to a coefficient and feature importance analysis.

MORS combines all user scores into a single learning process, which is the key contribution. This enables a thorough examination of the links between various airline service components and more precise assessments of the passenger experience. The study seeks to offer a solid foundation for comprehending passenger preferences and attitudes, while also showing telling relationships between all available scores.

In the next chapter, the details of the dataset along with the distributional issues and the preprocessing pipeline are presented. In the upcoming sections, the word embedding techniques and the selected baseline model architectures that this work exploits are analyzed. Finally, MORS will be described in depth by pointing out the major differences between Single and Multi-Output Regression. The figure 3.1 below depicts MORS in detail:

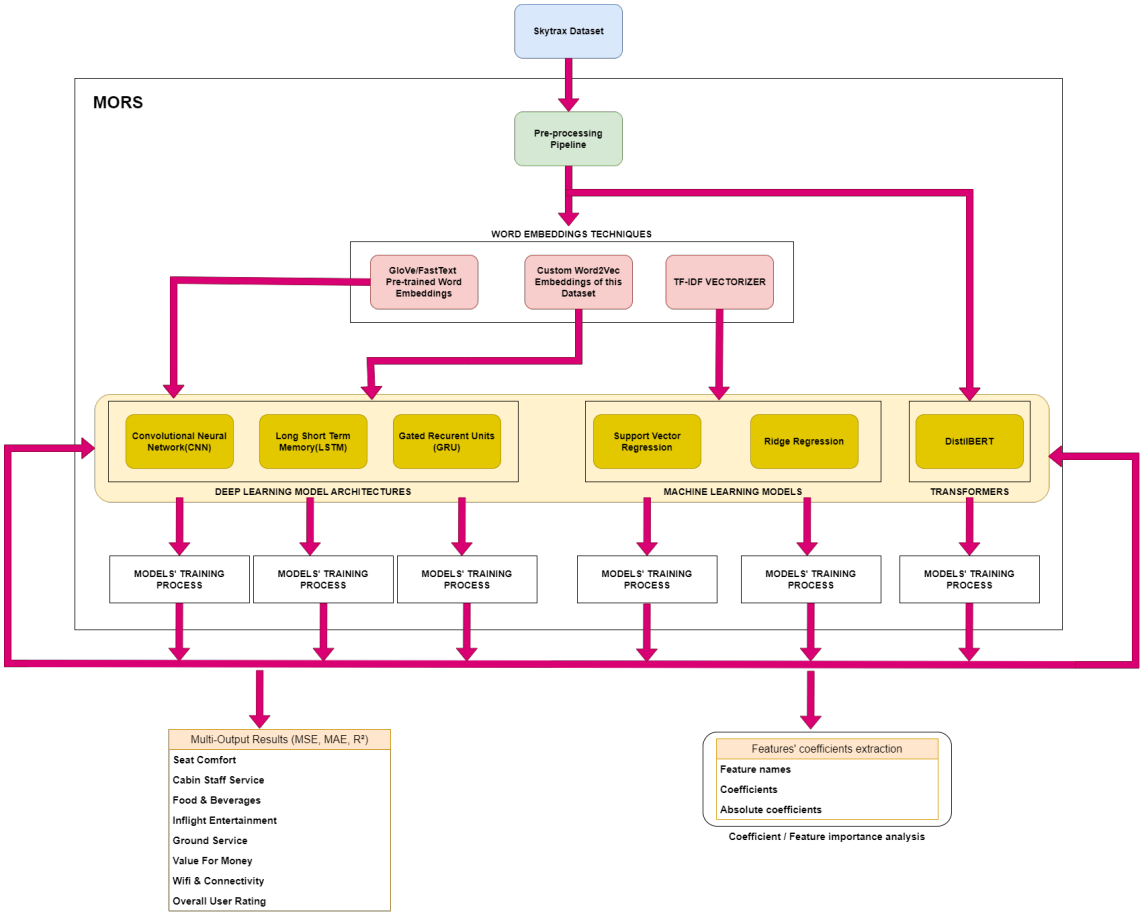


Figure 3.1: A detailed depiction of MORS

3.1 Word Embeddings

3.1.1 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) ("TF–IDF", 2010) is a vital weighting method in information retrieval and text mining. It assesses term importance using frequency within a document (TF)(Charatsaris, 2017) and rarity in the corpus (IDF). The tf-idf weight is:

$$tf_{n,d} \times \log\left(\frac{M}{\mathrm{df}_n}\right)$$

In this project, the TF-IDF method is effective since it converts text reviews into numerical representations while highlighting each term's importance inside a review and taking into account its rarity across the entire corpus. With this method, it is guaranteed that the feature vectors that are generated accurately represent the substance of each review.

There are various benefits of using the TF-IDF vectorizer with machine-learning regression models. By giving significant terms larger weights and frequent phrases lower weights, it improves the algorithms' capacity to focus on pertinent information, improving prediction accuracy and catching minor changes in passenger experiences.

Additionally, TF-IDF makes it possible to examine feature importance and coefficients in greater detail. It highlights the most influential elements on user scores by highlighting terms that are frequently used inside a given review but are relatively infrequent across the dataset, giving insightful information about important facets of airline service.

The capacity of TF-IDF to manage textual data's sparsity and large dimensionality is another advantage. It allows for the clear and informative depiction of extensive and varied text reviews while reducing computational complexity and maintaining essential information.

TF-IDF has limitations, though. It does not take into consideration word semantic relationships, which may be essential for comprehending the context of evaluations. Furthermore, it treats each term separately, ignoring crucial phrase-level information that can boost model performance.

In conclusion, using the TF-IDF ("TF–IDF", 2010) Vectorizer for our regression problem is a wise decision. By providing our machine learning regression models with discriminative and instructive feature representations, it improves prediction accuracy and provides a deeper understanding of the elements that influence passenger perceptions. Even though TF-IDF has many benefits, it is still important to use other word embedding approaches, since they can capture additional facets of the text input and enhance the models' overall performance.

3.1.2 GloVe

Natural language processing (NLP) jobs frequently use the highly efficient unsupervised learning algorithm known as **GloVe (Global Vectors for Word Representation)** (Pennington, Socher, & Manning, 2014). It tries to produce high-quality vector representations, also known as word embeddings, that accurately capture the semantic links between words based on how frequently those words occur together in a sizable corpus of text. The program accomplishes this by building a co-occurrence matrix (Charatsaris, 2017) that logs the frequency with which each word appears alongside each other throughout the corpus.

The vector representations of words are then calculated by GloVe (Pennington et al., 2014) by optimizing a global objective function based on the logarithm of the co-occurrence probability. The ability to dramatically improve the performance of deep learning models is one of the main advantages of using GloVe pre-trained word embeddings.
GloVe embeddings `https://towardsdatascience.com/light-on-math-ml-intuitive-guide-to-understanding-glove-embeddings-b13b4f19c010` have proven to be superb at capturing complex semantic connections between words, such as similarity or association. GloVe (Pennington et al., 2014) embeddings thus give deep learning models like CNN, GRU, and LSTM networks a contextualized representation of the input text. The models are better able to understand the reviews' underlying meaning thanks to this contextualization, which also improves accuracy and generalization for a variety of downstream tasks. GloVe (Pennington et al., 2014) embeddings also add to this study's investigation of feature importance and coefficients. GloVe embeddings enable the models to recognize the most important terms and phrases inside the reviews that have the biggest impact on the user scores by giving semantically enhanced word representations. This analysis provides an in-depth understanding of the elements that have a significant impact on passengers' perceptions and experiences.

Additionally, GloVe embeddings can successfully handle the sparsity and large dimensionality of textual data, resulting in more succinct and informative feature representations. GloVe embeddings record global word co-occurrence statistics. However, it is important to recognize that despite its advantages, GloVe might not completely capture the complexities of all review nuances and circumstances. It may not take into account domain-specific patterns because it is an unsupervised method, and it may generalize data across different domains. Additionally, the quantity and quality of the pre-training corpus have a significant impact on how well GloVe embeddings work. To guarantee that the embeddings are pertinent to this work's subject, great thought should be paid to the corpus selection.

3.1.3 FastText

FastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) is a well-known word embedding method made to handle the problems presented by words that are not commonly used and to record morphological information. It adds subword information to Word2Vec's skip-gram model, extending it. Each word is specifically represented as a group of character n-grams, which enables FastText to recognize internal word structure and handle unseen words by disassembling them into their component subwords. Furthermore, FastText has been successfully employed to build language models that achieve state-of-the-art results in text classification, sentiment analysis, and machine translation (Mouselimis, 2022).

When it comes to sentiment analysis and text regression for airline evaluations, FastText (Bojanowski et al., 2017) embeddings have two major advantages. First off, by including subword information, the model can capture more nuanced relationships between words, improving the model's ability to reflect context. This is especially helpful for jobs like sentiment analysis, where subtle changes in word choice can have a big impact on how people feel overall. Second, FastText's representation of words as a collection of character n-grams offers richer and more insightful insights when examining coefficients and feature importance. It enables a thorough analysis of the variables affecting customer attitudes and the elements that matter most in determining their evaluations.

For studying coefficients and feature importance, FastText word embeddings offer richer and more informative representations. FastText can determine the contribution of particular subword units to emotion or score prediction by presenting words as character n-grams.
This makes it possible to examine passenger opinion-influencing factors and the features that are most important in deciding ratings in depth. Additionally, FastText's accomplishments in NLP tasks and language models show its potency in identifying significant linguistic patterns, enhancing the generalization of deep learning models, and managing complexity in airline evaluations.

In conclusion, using FastText pre-trained word embeddings in deep learning models for sentiment analysis of airline passenger reviews improves the handling of words that are not part of a person's vocabulary, captures morphological information, and offers more illuminating representations for coefficient and feature importance analysis. These benefits improve the overall performance and interpretability of the models by enabling more precise and thorough insights into passenger feelings.

3.1.4 Custom Word2Vec

To learn distributed representations of words, a neural network model is trained on a domain-specific corpus(Skytax dataset) using the **Word2Vec** algorithm (Mikolov, Chen, Corrado, & Dean, 2013). This method seeks to record contextual and domain-specific details important for further analysis. **Continuous Bag of Words (CBOW)** and Skip-gram are the two main training techniques that Word2Vec makes use of.

1) **CBOW** (Mikolov, Chen, et al., 2013): Using the words in their immediate context, CBOW predicts the target word. The model reduces prediction error based on the context window to maximize word embeddings.

2) On the other hand, using a target word as a starting point, **Skip-gram** (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) predicts the words in the surrounding context. It is excellent at capturing rare words and performs well with short training datasets.

To lower prediction error during training, the neural network modifies word embeddings, resulting in representations that capture word semantic relationships. Due to the tendency of related words to have similar vector representations, Word2Vec facilitates tasks like word similarity and analogy reasoning. Word2Vec heavily relies on the idea of word context, presuming that words that appear in similar circumstances have comparable meanings. As a result, Word2Vec creates word embeddings from the learning context that capture semantic data and word relationships from the dataset.

When combined with deep learning models for natural language processing applications, Custom Word2Vec-trained word embeddings have many advantages. First off, Word2Vec can be trained on a domain-specific corpus to capture the particular nuances and contextual information pertinent to the target dataset. This is especially helpful for tasks like sentiment analysis in airline reviews, where domain-specific language and expressions can greatly influence the overall sentiment conveyed by passengers. Furthermore, Word2Vec's capacity to build embeddings that represent semantic connections between words makes tasks like word similarity and analogy reasoning easier to complete. This helps deep learning models better understand the connections between words and their meanings, which enhances performance in tasks requiring a nuanced understanding of language. Additionally, the effectiveness of the embeddings is greatly influenced by the concept of word context, which is fundamental to Word2Vec. This concept makes use of data about words that appear in comparable contexts to create embeddings that capture semantic information and word relationships, offering helpful insights into the underlying structure of the text data and revealing patterns and relationships that affect passenger opinions and scores.

Custom Word2Vec word embeddings are essential for coefficient and feature importance analysis because they make it possible to examine textual data in greater detail. Word2Vec, which represents words as distributed vectors, makes it possible to pinpoint the contributions of particular subword units to sentiment or score forecasts, giving granularity and insight into the variables affecting passenger perceptions and ratings. Additionally, these embeddings provide domain-specific context and semantic information to deep learning models, aiding language comprehension as well as sentiment analysis and score prediction tasks. The coefficient and feature importance analysis is improved by the embeddings' capacity to record fine-grained interactions between words, exposing the elements that affect passenger perceptions the most.

3.2 Baseline Models

3.2.1 Machine Learning Models

3.2.1.1 Support Vector Regression(SVR)

SVMs in regression aim to find a function that maps input features to the target variable `https://www.interviewquestionspdf.com/2023/08/45-capgemini-data-science-interview.html`. This function utilizes kernel functions to transform the features into a higher-dimensional space, a technique known as the "kernel trick" `https://medium.com/@stger040/a-beginners-guide-to-support-vector-machines-for-classification-1cb007e5c9d4`. This expansion helps to linearly separate the data in higher dimensions (Charatsaris, 2017). The optimal regression function is determined by minimizing a loss function that penalizes the distance between predicted and actual label values. This minimization is subject to a regularization term that controls the model's complexity. In this thesis project, text input data is converted into numerical features using sparse word embedding techniques. The linear kernel function is chosen as the kernel for the Support Vector Machine algorithm (*Support Vector Regressor SVR*, n.d.-a):

$$K(x, x') = x \cdot x'$$

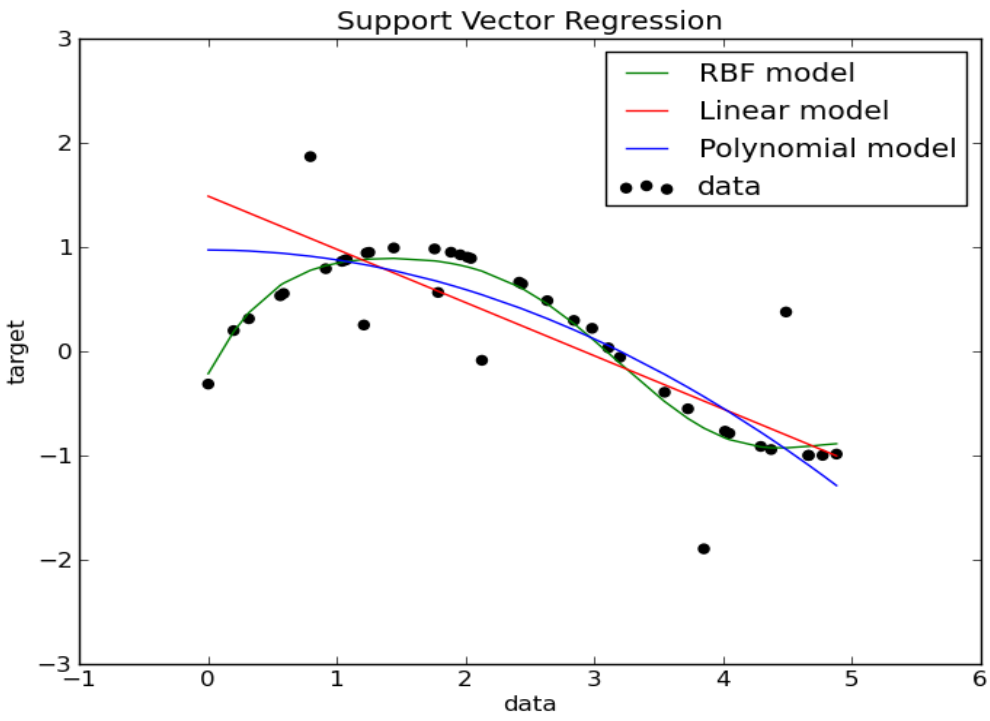


Figure 3.2: An illustration of a Support Vector Regression that depicts how different kernels including linear kernel fit the data on the graph .

To forecast pertinent scores in the context of airline reviews on the target dataset, the implemented method applies the Support Vector Regression (SVR) (*Support Vector Regressor SVR*,

n.d.-a) approach with a linear kernel function. The data is preprocessed before being supplied into the SVR model as numerical features created using the TF-IDF vectorizer.
The task benefits from the SVR algorithm's fast capture of complex patterns and nuances contained in text data, leading to enhanced score predictions. The SVR method is renowned for its effectiveness in handling linear connections. Additionally, the model's coefficient analysis offers insightful information on the roles that particular elements, like words or phrases, play in affecting passenger opinions. The SVR technique is a potent tool for sentiment analysis and feature importance analysis in airline reviews because it offers a deeper understanding of key aspects that affect passenger experiences when combined with TF-IDF-based feature representations.

3.2.1.2 Ridge Regression

Ridge Regression is a type of linear regression used for forecasting continuous target variables. It introduces a regularization term to the loss function that penalizes high coefficients in the regression equation, making the model smoother and more generalizable (Bishop, 2007; *Support Vector Regressor SVR*, n.d.-b). The regularization strength hyperparameter, represented by alpha, controls the penalty severity. The L2 norm regularization method is used to estimate the coefficients by shrinking them towards zero (Hastie, Tibshirani, & Friedman, 2001). This ensures better performance, especially when dealing with many input features.

The TF-IDF vectorizer is used in this work to turn textual data into numerical features, which are then used to forecast continuous target variables. There are various benefits to combining TF-IDF and Ridge Regression. First, TF-IDF captures the significance of words in each review, making textual data more illuminatingly represented. Ridge Regression can successfully simulate the link between input features and target scores because of its richer representation. Second, when working with high-dimensional text data, L2 norm regularization in Ridge Regression helps reduce overfitting and enhances the model's generalizability. Ridge Regression also offers an insightful examination of coefficients and feature relevance. The model highlights the most important terms or traits in predicting rating scores by estimating coefficients with regularization, enabling a more in-depth investigation of the elements that influence passenger opinions. This approach adds to improved performance in sentiment analysis and score prediction tasks by assisting in the knowledge of the critical factors that influence passenger evaluations.

3.2.2 Deep Learning Models

3.2.2.1 Convolutional Neural Network(CNN)

Even though Convolutional Neural Networks(CNNs) became fancy for analyzing visual images, plenty of works showed that they can be effective for many NLP tasks, e.g, semantic parsing (Yih, He, & Meek, 2014), search query retrieval (Shen, He, Gao, Deng, & Mesnil, 2014), sentence modeling (Kalchbrenner, Grefenstette, & Blunsom, 2014), text classification (Jacovi, Sar Shalom, & Goldberg, 2018) and text regression (Dereli & Saraclar, 2019) (Dereli & Saraclar, 2019).
The objective of this study is to propose a methodology for predicting relevant scores from customer text reviews by employing a Convolutional Neural Network (CNN) architecture integrated with word embeddings. Customer reviews are valuable sources of information, and accurately predicting relevant scores allows businesses to gain insights into customer experiences and satisfaction. The proposed approach leverages word embeddings to transform textual data into numerical sequences, enhancing CNN's ability to process and extract meaningful patterns effectively.

Given a batch size (batch_size = 32) and a sequence length (sequence_length = 300), the input sequence matrix X is obtained.
The embedding layer then maps each word index to a pre-trained word embedding vector, resulting in a tensor of shape (batch_size = 32, sequence_length = 300, embedding_dim = 300), where embedding_dim represents the dimensionality of the word embeddings. The CNN architecture comprises multiple Conv1D layers with varying kernel sizes K. The input sequences are convolved to extract local patterns. The output O_k of the k-th convolutional layer is calculated as $O_k = \text{ReLU}(X * W_k + b_k)$, where W_k denotes the weight matrix and b_k is the bias vector of the k-th convolutional layer.

Max pooling layers are then utilized to capture crucial features from the convolutional layer outputs. A receptive field of size P is chosen for each convolutional layer, selecting the maximum value and reducing the dimensionality of the features while preserving key information. Next, a fully connected layer with a ReLU activation function processes the concatenated tensor from the max pooling layers. Dropout regularization is applied to prevent overfitting. The output of this layer H is given by $H = \text{ReLU}(U * V + c)$, where U represents the concatenated tensor, V is the weight matrix of the fully connected layer, and c is the bias vector.

Finally, the output layer with a linear activation function predicts a continuous value for each relevant score. The proposed approach offers several benefits and advantages. Firstly, word embeddings enable the model to encode semantic meaning and relationships between words, facilitating a deeper understanding of customer reviews. Secondly, the CNN architecture with Conv1D layers allows the model to capture both detecting intricate details (local patterns) and broader trends (global patterns), enhancing the identification of complex relationships in the data. Moreover, the incorporation of max pooling layers emphasizes important features while reducing computational complexity, mitigating the risk of overfitting.

Furthermore, the utilization of word embeddings simplifies the coefficient/feature importance analysis. Since the embeddings already contain meaningful semantic information, the interpretation of coefficients representing feature importance becomes more transparent. This enables the identification of the most influential factors impacting customer experiences. In conclusion, the proposed CNN architecture utilizing the referred word embeddings is a promising and robust approach for predicting relevant scores from customer text reviews. The mathematical foundations and model interpretability provide valuable insights into customer experiences, facilitating evidence-based decision-making.

3.2.2.2 Long Short Term Memory(LSTM)

(Hochreiter & Schmidhuber, 1997) presented the Long Short-term Memory (LSTM) in 1997 as a solution to the problems with RNNs. Forget, update, and input gates are features of LSTM. The model's LSTM layers are built to identify long-term dependencies in the text data (Charatsaris, 2017). Let c_t stand for the cell state and h_t stand for the hidden state of the LSTM at time step t. The following is a definition of the LSTM update equations:

$$
\begin{aligned}
f_t &= \sigma(W_f \cdot [h_{t-1}, X_t]), \\
i_t &= \sigma(W_i \cdot [h_{t-1}, X_t]), \\
g_t &= \tanh(W_g \cdot [h_{t-1}, X_t]), \\
o_t &= \sigma(W_o \cdot [h_{t-1}, X_t]), \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}
$$

The forget gate, represented by f_t in these equations, selects how much data from the previous cell state should be discarded. The quantity of new information to be added to the cell state is controlled by the input gate i_t, and the potential new values are calculated by the output gate g_t. The information that will be output from the cell state is controlled by the output gate o_t. The LSTM layers effectively capture the sequential dependencies and allow the model to learn the complex patterns in the text reviews by using these updated equations.

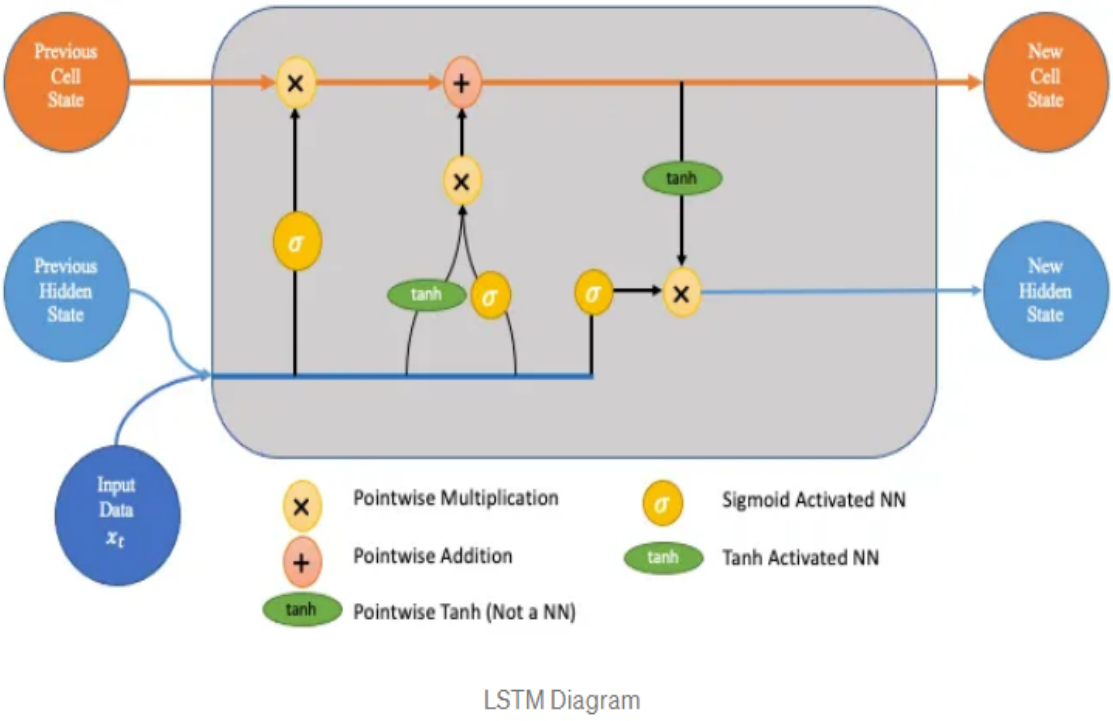


Figure 3.3: An LSTM diagram `https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9`.

The introduced model architecture of this work can be defined as follows:

- Input Sequences: The input data X comprises tokenized and padded sequences with a shape of (batch_size = 32, sequence_length = 300).
- Word Embeddings: An embedding layer transforms the input sequences into dense vectors of fixed sizes. Pre-trained word embeddings are stored in the embedding matrix E with a shape of (vocabulary size = 10000, embedding dimension = 300), where each row corresponds to a word in the vocabulary.
- Embedding Operation: The embedded input sequences X_{embedded} are obtained by performing the Embedding operation, which substitutes the relevant row of E for each word index in X.
- LSTM Layers: The LSTM layers capture temporal dependencies, the model employs two LSTM layers stacked on top of each other, with 128 and 64 units, both having a dropout rate of 0.2.
- Dropout Layer: The Dropout layer is applied after the LSTM layers, reducing overfitting by randomly setting some components to zero.
- Feature Extraction: Further feature extraction is facilitated by a fully linked layer with 64 units and ReLU activation.
- 2nd Dropout Layer: A second dropout layer is applied to prevent overfitting.
- Dense Layer: The last dense layer with a linear activation function projects the continuous value for each relevant score. The anticipated scores are represented by y_{pred}, computed as $y_{\text{pred}} = \text{Dense}(h_t)$.

The implementation of the LSTM-based architecture with word embeddings in this work brings several benefits to the regression task of predicting relevant scores from customer text reviews, as well as in the coefficient/feature importance analysis task :

- **Effective Sequential Data Handling:** The LSTM layers excel at capturing long-term dependencies in the text data, making them ideal for processing customer reviews that may have complex sequential patterns by effectively leveraging contextual nuances.
- **Meaningful Word Representations:** By utilizing pre-trained word embeddings, the model gains a better understanding of the words in the reviews, resulting in more accurate predictions and improved interpretability.
- **Improved Generalization:** The combination of stacked LSTM layers and dropout regularization enhances model generalization by mitigating overfitting, as dropout randomly deactivates neurons during training, preventing reliance on specific features and promoting the learning of more robust and adaptable representations from customer reviews.
- **Interpretable Feature Importance:** The LSTM-based architecture facilitates interpretable feature importance analysis by its ability to capture long-range dependencies, enabling identification of influential words through attention mechanisms, thus enhancing decision-making by offering transparent insights into the contribution of specific textual cues to relevant score predictions.
- **Holistic Customer Experience Analysis:** By mapping text reviews to each relevant score, the LSTM-based model empowers comprehensive study and evaluation of customer experiences, assisting businesses in identifying strengths and weaknesses.

3.2.2.3 Gated Reccurent Unit(GRU)

GRU (Gated Recurrent Unit), which was introduced by (Cho et al., 2014), seeks to address the vanishing gradient issue that arises with a conventional recurrent neural network. GRU can also be viewed as an adaptation of the LSTM (Chung, Gulcehre, Cho, & Bengio, 2014) due to the similarities in their designs and, in some situations, they are equally impressive in the results they produce.

One definition of the GRU update equations is:

$$
\begin{aligned}
z_t &= \sigma(W_z \cdot [h_{t-1}, X_t]) \\
r_t &= \sigma(W_r \cdot [h_{t-1}, X_t]) \\
c_t &= \tanh(W_c \cdot [r_t \odot h_{t-1}, X_t]) \\
h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot c_t
\end{aligned}
$$

where W_z, W_r, and W_c are weight matrices, *sigma* stands for the sigmoid activation function, and *odot* signifies element-wise multiplication.

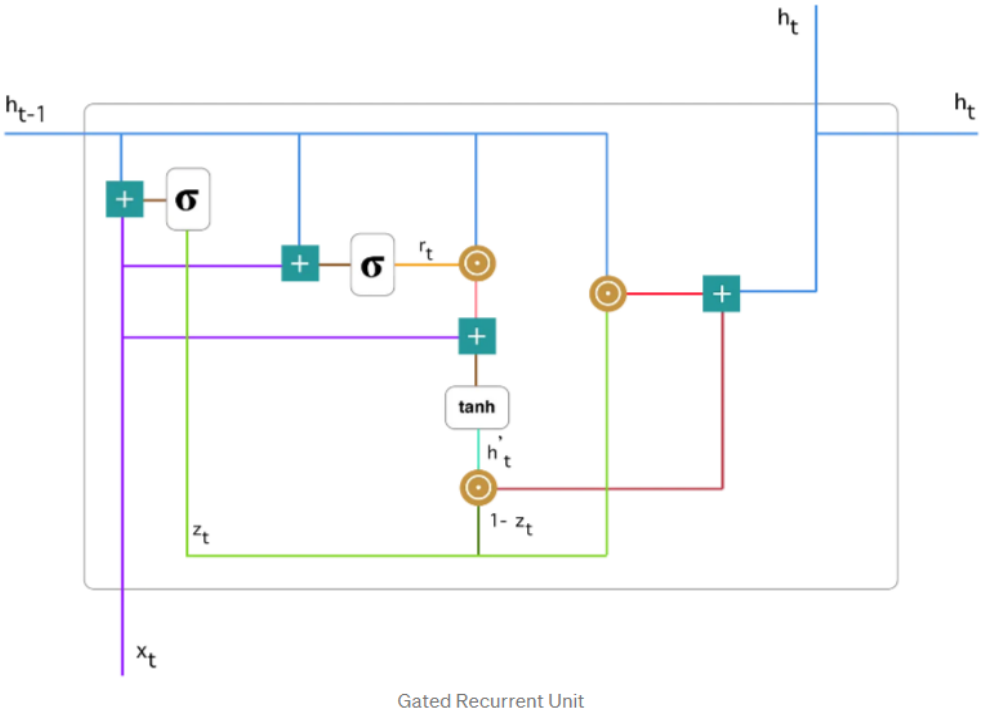


Figure 3.4: A detailed version of a single Gated Recurrent Unit (GRU) `https://towardsdatascience.com/understanding-gru-networks-2ef37df6c9be`.

The introduced model architecture of this work can be defined as follows:

- Input Sequences: The input data X comprises tokenized and padded sequences with a shape of (batch_size = 32, sequence_length = 300).

- Word Embeddings: An embedding layer transforms the input sequences into dense vectors of fixed sizes. Pre-trained word embeddings are stored in the embedding matrix E with a shape of (vocabulary size = 10000, embedding dimension = 300), where each row corresponds to a word in the vocabulary.
- Embedding Operation: The embedded input sequences X_{embedded} are obtained by performing the Embedding operation, which substitutes the relevant row of E for each word index in X.
- GRU Layers: The GRU layers capture the sequential dependencies in the text data. At each time step t, the GRU's hidden state is denoted as h_t.
- Dense Layer: The last dense layer with a linear activation function projects the continuous value for each relevant score. The anticipated scores are represented by y_{pred}, computed as $y_{\text{pred}} = \text{Dense}(h_t)$.

The implementation of the GRU-based architecture with word embeddings in this work brings several benefits to the regression task of predicting relevant scores from customer text reviews as well as in the coefficient/feature importance analysis task :

- **Efficient Handling of Sequential Data:** GRU's gating mechanisms enable efficient processing of sequential data such as customer reviews by selectively retaining relevant information, achieved through the modulation of information flow, which aids in capturing dependencies and patterns within the text while preventing vanishing gradient issues.
- **Utilization of Word Embeddings:** Pre-trained word embeddings provide meaningful word representations, leading to better understanding and more accurate predictions.
- **Interpretability through Feature Importance:** GRU's implementation facilitates interpretability by enabling feature importance analysis through its ability to capture sequential dependencies, which empowers the identification of influential words in determining relevant scores through attention mechanisms, thereby enhancing model transparency and decision-making.
- **Generalization to Different Reviews:** The model's adeptness in generalizing to diverse customer reviews stems from its ability to learn versatile and abstract representations by capturing meaningful sequential patterns, which enables it to effectively adapt to different review contexts and enhance its practical usability.
- **Robustness to Noisy Data:** GRU's gating mechanisms enhance robustness to noisy data by enabling selective information retention, which prevents undue reliance on noise and contributes to reduced overfitting, thus improving the model's capacity to learn relevant patterns from the data.

3.2.3 Transformers

3.2.3.1 DistilBERT

Due to the need of having a less massive model than BERT for the sake of this thesis's experiments, DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2019)architecture was used which is a compression of the original BERT (Devlin, Chang, Lee, & Toutanova, 2018) model, by knowledge distillation technique in which a small model is trained to reproduce the behavior of a large model.

A transformer-based model called DistilBERT has been pre-trained using a sizable corpus of text data. The DistilBERT tokenizer is used to first tokenize the input texts, turning them into a list of numeric tokens. Let $X_{\text{train\_tokens['input\_ids']}}$ represent the tokenized input IDs and $X_{\text{train\_tokens['attention\_mask']}}$ represent the attention masks.

The DistilBERT model (Sanh et al., 2019), which consists of several transformer layers, is then fed the tokenized inputs. Each transformer layer uses self-attention processes to take note of the relationships between words and other context-relevant information in the text. A series of contextualized embeddings, designated as E, is the DistilBERT model's output.

The embeddings E are subjected to a global average pooling procedure to provide a fixed-length representation of the text. This pooling procedure generates a fixed-length vector P by averaging the embeddings across the sequence dimension. A dense layer with linear activation receives the vector P as input and predicts a continuous value for the selected label score.

Below are some of the most significant advantages that this architecture brings to the regression task of predicting the relevant scores from customer text reviews and the coefficient/feature importance analysis:

- **Contextualized Text Representation:** DistilBERT's contextualized text representation is achieved through transformer-based attention mechanisms, allowing it to understand the interdependencies of words within sentences, thus improving its regression pattern learning by considering intricate linguistic relationships and contextual cues.
- **Efficient Processing:** DistilBERT's efficiency in tokenizing and processing large-scale datasets is attributed to its transformer architecture, which learns hierarchical embeddings from raw text, eliminating the need for labor-intensive feature engineering and enables automated extraction of informative representations directly from the data.
- **Feature Importance Analysis:** Through attention mechanisms and fine-tuned contextual embeddings, DistilBERT facilitates feature importance analysis by assigning varying weights to tokens, highlighting their significance in predicting user ratings based on their contribution to the overall context, thus enabling transparent insights into the influential linguistic elements.
- **Quantitative Insights:** DistilBERT's coefficient analysis provides quantitative insights by assigning numerical values to token contributions, facilitating informed decisions through the precise measurement of each token's impact on predicting user ratings, and enhancing the interpretability of the model's results.
- **Automated Ranking:** DistilBERT's automated token ranking expedites the identification of key predictive factors by assigning tokens relative importance scores through attention mechanisms, streamlining the process of pinpointing significant contributors to predictions within the context of user ratings.
- **Enhanced Transparency:** Feature importance analysis within DistilBERT fosters transparency and trust in predictions by revealing the quantitative influence of each token, establishing a clear link between linguistic cues and model outputs, thereby enhancing the model's interpretability and its credibility in decision-making.

- **Domain-Specific Insights:** DistilBERT's domain-specific insights stem from its pre-training on vast text corpora, allowing it to recognize and leverage domain-specific language patterns, which enables the extraction of nuanced insights tailored to the particular domain, thus contributing to improved user satisfaction.
- **Iterative Refinement:** DistilBERT's insights facilitate iterative model refinement by pinpointing areas for improvement through the analysis of token contributions, guiding the adjustment of model parameters or architecture, leading to enhanced performance in predicting user ratings.
- **Generalizability:** DistilBERT's generalizability to diverse regression tasks across domains arises from its capacity to learn universal linguistic patterns during pretraining, allowing it to adapt and fine-tune for specific tasks, thus demonstrating robust predictive capabilities across a spectrum of contexts.

3.3 Multi-Output Regression System (MORS)

The following section gives an overview of how Multi-Output Regression works. Furthermore, the design of the system will be thoroughly examined on how MORS utilizes the combination of the above elements of the system 3.1 that were discussed in the previous sections of this chapter, with special attention paid to outlining the key differences between single and multi-output regression.

3.3.1 Multi-Output Regression Overview

In the Multi-Output cases `https://towardsdatascience.com/machine-learning-on-multioutput-datasets-a-quick-guide-ebeba81b97d1`, there is more than one target column than the classic machine learning tasks, and the goal is to train a model capable of predicting every one of them at the same time. There are three major types of Multi-Output Tasks:

- Multilabel
- Multiclass-Multioutput
- Multi-Output Regression

Multiple numerical attributes are predicted for each sample using Multi-Output Regression. Each property is a numerical variable, and there must be at least two anticipated properties for each sample. Using information gathered at a certain place, for instance, one may anticipate both wind speed and wind direction, both expressed in degrees. Each sample would have information collected at a single place, and the output for each sample would include the wind speed and direction. For the needs of this thesis, Multi-Output Regression is exploited to create MORS for the Skytrax dataset that this work is experimenting on.

3.3.2 System Architecture

This work introduces **Multi-Output Regression (MORS)**, an advanced technique that surpasses traditional single-output regression paradigms. MORS has the distinctive capability of simultaneously predicting eight distinct target scores, as illustrated in Figure 3.1. The architecture of MORS builds upon the selected multi-layered baseline models that harness the potency of contextual word embeddings, as previously mentioned.

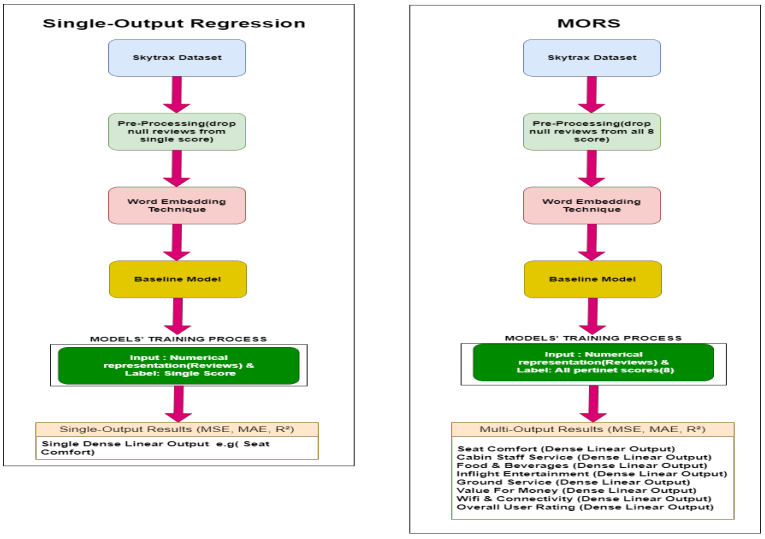


Figure 3.5: Single vs Multi-Output Regression

Single-Output vs. Multi-Output Regression

Single-Output Regression Process

In the context of Single-Output Regression, the data pre-processing phase 4 takes place. Tokenization(TF-IDF Vectorizer, Keras Text Pre-processing Tokenizer, DistilBERT Tokenizer) and pre-trained word embeddings are then applied to the pre-processed text data utilizing the word embedding techniques mentioned in the previous sections, to create a numerical representation of words(excluding the Machine Learning models). Each of the selected model architectures comprises its layers that were discussed on 3 and the final output layer consists of a single dense unit with linear activation for predicting the "selected score" target variable. The selected model is compiled with a loss function and trained on the training data. The evaluation of the model involves the metrics that this thesis uses(MSE, MAE& R^2), providing insights into the model's performance.

Multi-Output Regression Process

In contrast, Multi-Output Regression introduces an extension to the Single-Output approach. The data preprocessing remains consistent, including tokenization and word embedding tasks, apart from the fact that every review that contains a single null score of all the scores is removed. However, the model architecture features multiple dense output layers, each dedicated to predicting a distinct target variable e.g.('Seat Comfort', 'Cabin Staff Service', etc.). This allows the model to simultaneously predict multiple aspects of the input data. While compiling and training the model, distinct loss functions are assigned to each output, ensuring the model's optimization for multiple tasks. The evaluation process now includes metrics specific to each target variable, offering a nuanced understanding of the model's performance on different aspects. Overall, Multi-Output Regression extends the capabilities of Single-Output Regression by enabling the prediction of multiple related variables in a coordinated manner.

Diverging from single-output regression, where a solitary target variable(airline score) is forecasted, MORS adopts a pioneering approach. By incorporating a series of densely connected linear output layers, MORS constructs a framework to predict multiple target variables concurrently. This approach underpins MORS' ability to holistically comprehend customer experiences, amalgamating all user scores from Skytrax's dataset into a unified learning process (Borchani, Varando, Bielza, & Larranaga, 2015). As a result, the relationships between diverse airline service components are scrutinized more effectively.

Furthermore, MORS exploits the inherent linkages between user scores (Alvarez & Lawrence, 2008). This unique feature empowers the models to leverage interdependencies and shared knowledge, leading to joint predictions and reduced error. Notably, MORS is adept at identifying elements with simultaneous impacts on multiple user ratings. Coefficient and feature importance analysis across various output dimensions, as shown in Figure 3.1, reveals pivotal factors that shape the overall customer experience. This insight equips airlines to enhance customer happiness and loyalty through targeted improvements.

Of equal significance, MORS enhances model interpretation (Borchani et al., 2015). By jointly predicting multiple ratings, it provides a comprehensive understanding of factors influencing the complete customer experience. This transparency facilitates informed decision-making, empowering stakeholders with a deep comprehension of the underlying mechanisms guiding the predictive outcomes.

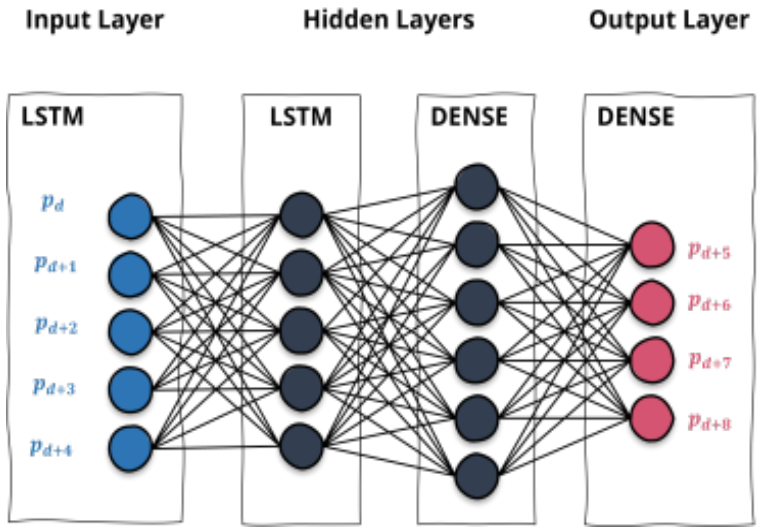


Figure 3.6: An LSTM model architecture with Multi-Output Regression example.

Chapter 4

Data

The following section discusses the details of the dataset used for this master thesis. The first section describes the dataset and the challenges that come with it, the second chapter presents the pre-processing pipeline for all model categories that were used and tested in this work. Approximately 108,000 airline reviews from SkyTrax (`https://www.airlinequality.com`) were used as the dataset for this thesis research. The UK-based international air transport rating agency SkyTrax is committed to improving the passenger experience at airports and airlines all over the world. Users can obtain condensed user evaluations, ratings, and information about overall service performance on the SkyTrax website. The dataset consists of individual reviews that provide various airline experience-related details. These reviews offer insightful information on a variety of topics, including onboard amenities, customer service, flight experiences, etc. It is important to note that the dataset includes evaluations from a wide range of internationally operating airlines, providing a thorough insight into the airline sector. It is also worth noting here the work of (Charatsaris, 2017) who annotated manually the actual relevant scores of some of the aspects that are mentioned above and created a stable pre-processing pipeline for the data. This thesis (Charatsaris, 2017) will serve as a guiding framework, and its steps and methodologies will play a pivotal role in shaping and informing my research.

The dataset's complete contents are listed below:

1. Name of reviewer.
2. Country of reviewer.
3. Airline name.
4. Section being reviewed.
5. Date of review.
6. Overall User Rating.
7. Title of review.
8. Verified trip.
9. Review count of reviewer.
10. Actual review (text).

11. Date of response.
12. Actual response (text).
13. Type of traveler.
14. Type of seat.
15. Actual route.
16. Date flew.
17. Seat comfort score.
18. Cabin staff service score.
19. Food & beverages score.
20. In-flight entertainment score.
21. Ground service score.
22. Wi-fi & Connectivity score.
23. Value for money experience score.
24. Recommendation decision.

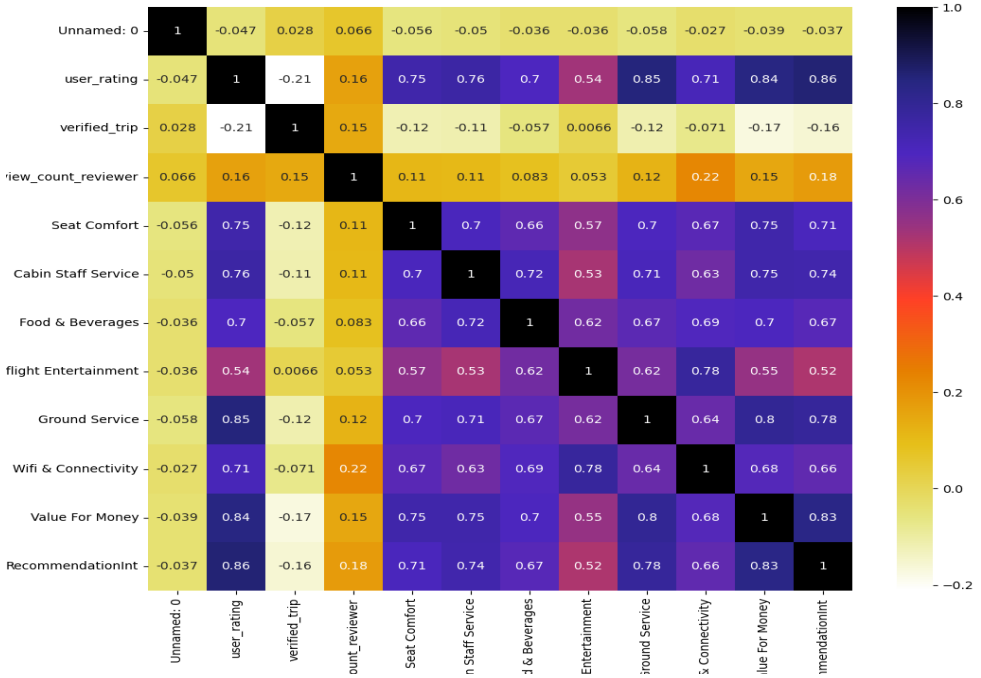


Figure 4.1: HeatMap of the Pearson Correlation in the contents of the dataset used for this project, that demonstrates 60% or higher correlation between most of the scores.

The scores that are immediately significant to the thesis project's objective span a wide spectrum, including user ratings 4.2. The remaining scores fall between the ranges of 0 and 5, with 0 denoting the lowest score and 5 denoting the highest. Higher scores on this scale imply better performance, allowing for the measurement of performance along a continuum. The user rating values vary from 1 to 10, making it possible to make a more complex assessment. The heatmap of the Pearson correlation was calculated and it is illustrated on 4.1. Having confirmed that all relevant scores of the Skytrax dataset that this thesis utilizes have a high correlation between them, the motivation of combining all scores using MORS rose since each factor(important feature/keyword) of a specific score might be commonly describing more than one score leading improving the performance of the system.

4.1 Insufficience & Challenges

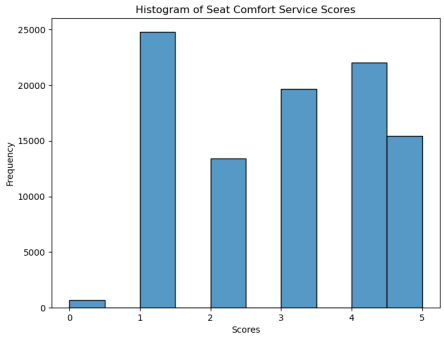
The dataset contains reviews, but several columns have a disproportionately high amount of null values, which may have an impact on how prediction models for the target scores are trained. In particular, lots of review's assigned scores are suffering from being empty and the quantity of those is depicted in the below matrix:

Score Type	Empty cells
Seat Comfort	11111
Cabin Staff Service	11277
Food & Beverages	25369
In-flight & Entertainment	38846
Ground Service	40360
Wi-fi & Connectivity	87187
Value For Money	1813
User Ratings	858

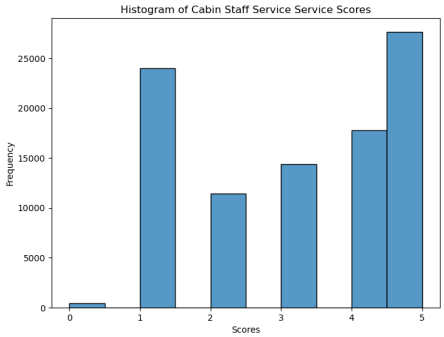
To guarantee the dependability and correctness of the models, it is essential to treat these missing variables and the uneven distribution of scores effectively during data preparation. Illustration in figures 4.2.

Several significant observations can be drawn from the histograms and score distributions. The distribution of the results for Seat Comfort and Cabin Staff Service is fairly balanced across the whole rating scale. However, there is a noticeable disparity between the results for Food & Beverages, In-Flight Entertainment, Ground Service, Wi-Fi & Connectivity, and Value For Money. Particularly, score 1 is vastly overrepresented and outscores the other scores by a wide margin. As these scores contain six label classes (0-5) but display an unequal distribution of text evaluations, this imbalance presents difficulties and challenges for the learning and inference processes of the models that will be used in this work. The absence and the pretty low amount of assigned scores of 0 in the "Ground Service" & Wi-Fi Connectivity and "Value For Money" categories respectively, should also be emphasized because this can affect how the model learns, as all models used in the experiments would only attempt to predict five labels rather than the actual six labels.

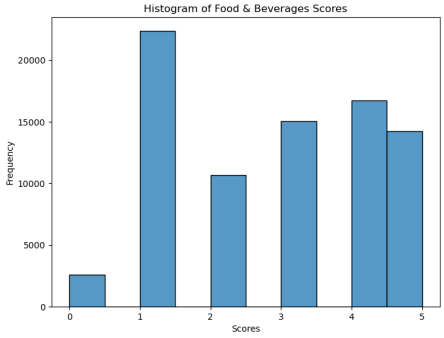
Additionally, there is a noticeable imbalance in the User Rating scores, which range from 1 to 10 and are given by all people in the reviews (apart from 858 cases). The score value 1 is notable because it is given around 20,000 times more often than the second-highest score value, which is 10. The learning process and inference outcomes of the models are significantly impacted by this



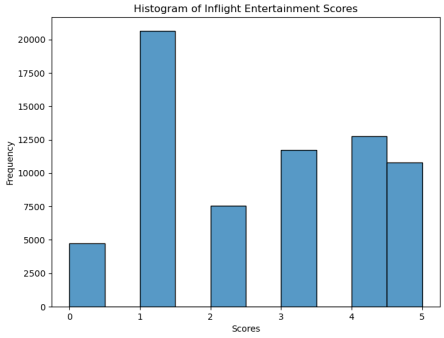
(a) Seat Comfort Scores Histogram



(b) Cabin Staff Service Scores Histogram

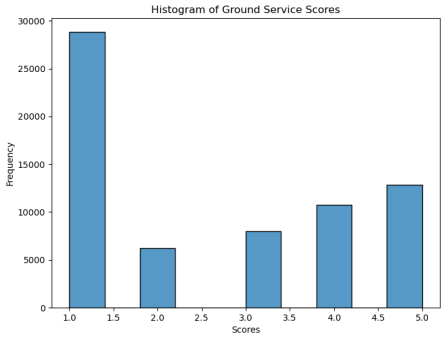


(c) Food & Beverages Scores Histogram

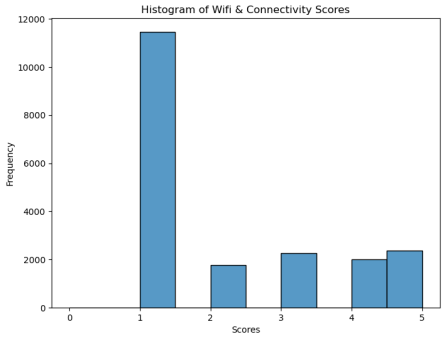


(d) Inflight Entertainment Scores Histogram

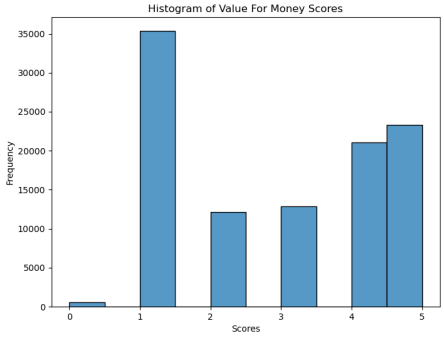
Figure 4.2: Page 1 of 2 with scores distribution histograms.



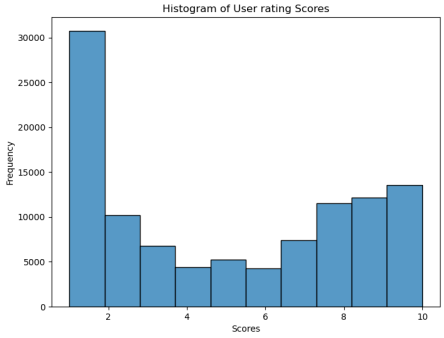
(e) Ground Service Scores Histogram



(f) Wifi & Connectivity Scores Histogram



(g) Value For Money Scores Histogram



(h) Overall User Rating Scores Histogram

Figure 4.2: Page 2 of 2 with scores distribution histograms.

significant variance in the distribution of user ratings. As a result, this particular field consistently produces the highest mean absolute error across all studies, illustrating how challenging it is to anticipate user ratings with any degree of accuracy in comparison to other scores.

4.1.1 Inefficient Scores Issues

One of the main focuses of this master's thesis is to address the issue of biased aspect scores within the dataset. A notable challenge arises from the presence of numerous reviews that do not mention certain aspects, yet these aspects are assigned low scores. For instance, consider the following review: "Please do yourself a favor and do not fly with Adria. On the route from Munich to Pristina in July 2019, they lost my luggage and for 10 days in a row, despite numerous phone calls, they were not able to locate it. 11 days later, the luggage arrived at the destination."

While this review is marked as "not verified," it is observed that seat comfort, food & beverages, cabin staff service, and Wi-Fi & connectivity receive extremely low or no scores at all, despite the review not referring to any of these aspects. This dataset's instability poses several challenges in accurately predicting the overall rating. Addressing this issue is crucial to improve the performance and reliability of the prediction models.

4.2 Pre-Processing Pipeline

To ensure data cleanliness and uniformity, several procedures are used to preprocess the text reviews. Firstly, the reviews that contain null scores are not considered for further pre-processing. To ensure consistency in text casing, the reviews are first transformed to lowercase using the `lower()` function. Using the `replace()` function and a regular expression pattern, special characters are then eliminated by replacing URLs and mention with empty strings. Using the `re.sub()` method and a pattern that matches alphanumeric characters with digits, and words containing digits are removed.

The `remove_punctuations()` method is used to remove punctuation marks. The NLTK library's `stopwords.words('english')` collection is also used to filter out commonly used English stop words, with the outcome being stored in the `stop_words` column. The `decontracted()` method is used to expand contractions in the reviews by substituting their expanded variants for contracted ones. For many reasons, the preprocessing pipeline does not include lemmatization, a typical method for word normalization. First off, keeping the original word forms preserves the dataset's context and subtle meaning fluctuations. The models can capture particular nuances and contextual information that might be lost when words are reduced to their simplest form by avoiding lemmatization. Additionally, the models can acquire more exact links between words and the goal values, since they have access to a bigger lexicon without lemmatization. The models can identify more intricate patterns because the fine-grained differences between words are preserved. Additionally, certain word forms or specific task-related elements may have a big impact when it comes to forecasting all existing scores. Lemmatization may obscure these important patterns by condensing many word forms into a single lemma. The models can capture the distinctive linguistic patterns associated with the e.g. (seat comfort score) by not applying lemmatization and instead making use of the original word forms. Last but not least, taking into account the characteristics of the dataset, lemmatization may not be as effective if it comprises informal or colloquial language, particular jargon, or domain-specific terminology and that is why every experiment reduced the performance of all models.

Chapter 5

Experimental Evaluation

The handling of the data, the metrics used, the models & hyperparameters of the experiments, as well as the technical setup of the hardware that the experiments were conducted, will all be examined in this section to characterize the experimental setup. The de facto standard for all tests are the airline reviews from SkyTrax (`https://www.airlinequality.com`), which were explained in Chapter4. Finally, the obtained experiment results are demonstrated in the form of arrays and intuitive analysis of the metrics and the coefficient/feature importance analysis takes part in the end of this chapter.

5.1 Experimental Setup

The careful design of the experimental setting is crucial to the validity and reliability of our study. Starting with the complex pre-processing of the raw data provided in Chapter4, this crucial stage explains the wide range of components concerning the handling of the data in each experiment. The metrics employed for this task are discussed with meticulous consideration, and chosen to serve as compasses guiding us toward accurate and meaningful predictions. This investigation of emotion within airline evaluations is both granular and thorough by carefully choosing the proper models that are complemented by hyperparameters which are calibrated to maintain the delicate equilibrium between model complexity and generalizability. The technical setup for this extensive empirical pursuit is revealed as the scene is being established. Finally, this holistic design ensures that the subsequent experiment results carry not just insights, but also the weight of methodological rigor, rendering our findings both impactful and trustworthy.

5.1.1 Data Set

As explained in Chapter 4, the SkyTrax platform's `https://www.airlinequality.com` extensive and detailed landscape of airline reviews serves as the starting point for this set of experiments. The same pre-processing pipeline is followed for all the experiments, as stated in Chapter4. In the context of training models to predict individual scores, this work adopts a tailored approach. Particularly, when training a model to predict a single score, removal of data instances with null values (see here 4.1) in the corresponding score column takes place by this command: df.dropna(subset=['chosen_score']), which eventually leads to a different amount of training data for every score.

However, when transitioning to MORS in which the models aim to predict all scores simultaneously, a more comprehensive approach is undertaken. The process of df.dropna(subset=[”]) is executed for all score columns. The purpose of this expanded breadth of data removal, even though it results in more aggressive data removal(MORS trains on approximately 18000 reviews) and a potentially smaller dataset, is to establish a uniform and consistent training environment for all multi-output models.
The strategic trade-off between data volume and multi-output learning appropriateness is a key component of this data handling methodology. Even though the experiments are not fully fair and there are limitations and future work for this dataset that are discussed in detail in Chapter 6, MORS outperforms all single prediction models and shows the potential for much more performance improvement if the score labels distribution problem of the dataset is handled appropriately.

5.1.2 Metrics Employed

A variety of metrics are used in this study's evaluation and analysis of the experimental data to gauge how well the constructed models predict outcomes. These measures are essential tools for determining how accurate and reliable the predictions made by the models are. As the main evaluation metrics, Mean Squared Error (MSE), Mean Absolute Error (MAE), and R Squared (R^2) are used. These metrics are crucial in assessing the overall effectiveness of the models across various target variables because they provide insightful information about the degree of variation between the projected and actual values.

These metrics are routinely calculated and examined as part of the fine-tuning and validation processes for the models, allowing for a robust evaluation of the model's accuracy in predicting different user experience ratings. By using these measures, a comprehensive grasp of the models' prediction abilities and constraints is attained, aiding in the thorough evaluation of the research findings.

5.1.3 Models & Hyperparameters

The hyper-parameters and the training settings that were used in these experiments are mentioned below :

Machine Learning Models

- Baseline Models Architectures : SVR, Ridge
- Word Embedding Techniques : TF-IDF
- alpha : 1.0
- test size : 20%
- random state : 42

Deep Learning Models

- Baseline Models Architectures : CNN, GRU, LSTM
- Word Embedding Techniques : GloVe(pre-trained 300d), FastText(pre-trained 300d), Custom(Word2Vec 300d)
- test size : 20%

- random state : 42
- Batch size : 32
- Epochs : 10
- Learning Rate : 0.001(default)
- Dataset Input Shape : (300,)

Transformers

- Baseline Models Architectures : pre-trained DistilBERT('distilbert-base-uncased')
- Tokenization : pre-trained DistilBERT Tokenizer('distilbert-base-uncased')
- test size : 20%
- random state : 42
- Batch size : 32
- Epochs : 10
- Learning Rate : 0.001(default)
- Dataset Input Shape : (300,)

5.1.4 Technical Configuration

For the carried out experiments, to complete the training/evaluation process of all the models used, the Google Colab Pro platform was utilized along with a GPU accelerator of T4 Tesla(3.6 billion transistors, 16GB GDDR6 memory), a CPU of Intel Xeon with 2 vCPUs(virtual CPUs) and 25.5GB of RAM.

5.2 Experiment Results

In this section, the experiment results are reviewed for the performance of all baseline models with one linear Dense layer that predicts a single score at a time compared to MORS performance. The first part describes the choice of selecting MAE as the main metric of these experiments and why it is considered the most stable out of the rest. Afterward, the depiction of the comparison of the results takes place along with intuitive insights for all tested models. Finally, coefficient/feature importance analysis takes place by showing the explainable dimension of the most important facts/words that are shared across both the single learning baseline models and MORS.

5.2.1 Mean Absolute Error

The experiments that are shown in the next subsection are carried out by only illustrating MAE as the main metric since out of the rest is considered the most stable metric. The list below provides an analytical explanation of why this work chose MAE for its main metric :

- **Interpretability:** The average absolute difference between projected values and actual values is directly measured by MAE. Because it shows the average error magnitude in the same unit as the original data, it is easier to understand. MSE in contrast, square the mistakes, which can occasionally make the results more difficult to understand.

- **Robustness to Outliers:** Compared to MSE and R^2, MAE is less susceptible to outliers. In MSE and R^2 computations, outliers have a squared influence, which could result in skewed findings. When dealing with data points that greatly depart from the norm, MAE's robustness increases since it regards all mistakes as linear functions.
- **Real-world Relevance:** When represented in the data's native units, a prediction error can often be quantified and understood more clearly in real-world situations. Errors are interpreted in practice in a manner that is well-aligned with MAE's clear interpretation in the context of the original data units.
- **Penalization of Larger Errors:** Due to the squaring effect, MSE gives larger errors higher weights, whereas MAE treats all errors equally. When avoiding disproportionately high penalties for larger deviations is desirable or when larger errors are not necessarily more significant, it can be advantageous to do this.
- **Stability:** MAE is considered to be a more stable metric than R^2 & MSE. When working with datasets of different scales or comparing models across datasets, this stability is very crucial. In certain situations, R^2 tends to magnify model differences, which could not correctly reflect the performance of the model.
- **Model Assessment:** The average error magnitude can be shown clearly thanks to MAE. This is particularly helpful for comparing models and evaluating their general performance. While outliers and different scales may have an impact on R^2, MAE provides a more balanced picture.

Lastly, the full set of experiments that include all three chosen metrics are analytically illustrated in the Appendix 7 of this thesis project.

5.2.2 Machine Learning Models

5.2.2.1 TF-IDF Vectorizer

Model	SC	CSS	FB	GS	IFE	VFM	Wi-Fi	UR
SVR	0.77	0.80	0.81	0.73	0.93	0.70	0.80	1.40
SVR(MORS)	0.77	0.86	**0.77**	0.73	**0.81**	0.70	0.79	**1.22**
Ridge	0.78	0.85	0.85	0.77	0.95	0.75	0.84	1.42
Ridge(MORS)	0.79	0.86	**0.78**	**0.74**	**0.83**	**0.71**	0.84	**1.25**

While the dataset exhibits a substantial volume of reviews, the presence of null scores poses limitations on the available training data and introduces disparities between Single-Output and Multi-Output Regression approaches. Moreover, the inherent sparsity within text data, where a multitude of words is coupled with a minority of terms that pertain to distinct categories, contributes to the challenges faced in predictive modeling. The inherent high dimensionality of the TF-IDF representation exacerbates this issue by affecting the model's capacity to discern meaningful patterns. As such, it becomes imperative to explore techniques that can navigate this intricate landscape effectively. Additionally, the paramount importance of data explainability is highlighted through TF-IDF's propensity to emphasize common words across diverse categories. However, this emphasis inadvertently diminishes feature diversity within each category, exemplified by the recurrent occurrence of terms such as 'thank,' 'excellent,' 'great,' 'worst,' and 'not.' This underscores the limitations of relying solely on ubiquitous terms to capture nuanced insights specific to each aspect.

While larger datasets indeed offer advantages, such benefits tend to wane when they come at the cost of diluted and sparse information. Focusing on the MORS approach, the machine learning models manage to strike a slightly better balance by curating more pertinent and focused data for each category, thereby preserving informational quality while optimizing for quantity. In conclusion, the strategic adoption of MORS, coupled with these machine learning models, reveals performance enhancements, as evidenced by the MAE metric. MORS consistently outperforms its single-output counterparts, capitalizing on its capacity to jointly model multiple aspects. This achievement demonstrates the merit of a collective approach in tackling the intricacies of aspect prediction within the context of airline passenger reviews.

5.2.3 Deep Learning Models

5.2.3.1 GloVe Pre-Trained Word Embeddings

Model	SC	CSS	FB	GS	IFE	VFM	Wi-Fi	UR
CNN	0.77	0.82	0.82	0.68	0.92	0.71	0.78	1.37
CNN(MORS)	0.76	**0.54**	0.80	**0.55**	**0.70**	**0.59**	**0.66**	**0.57**
GRU	0.75	0.73	0.77	0.66	0.88	0.65	0.76	1.15
GRU(MORS)	**0.72**	**0.58**	0.77	**0.56**	**0.70**	**0.59**	**0.63**	**0.63**
LSTM	0.76	0.72	0.73	0.63	0.84	0.65	0.75	0.96
LSTM(MORS)	0.76	**0.55**	0.83	**0.54**	**0.71**	**0.59**	**0.65**	**0.60**

5.2.3.2 FastText Pre-Trained Word Embeddings

Model	SC	CSS	FB	GS	IFE	VFM	Wi-Fi	UR
CNN	0.76	0.80	0.83	0.71	0.99	0.72	0.81	1.32
CNN(MORS)	0.76	**0.54**	0.83	**0.53**	**0.71**	**0.58**	**0.67**	**0.61**
GRU	0.71	0.72	0.74	0.60	0.85	0.64	0.76	1.15
GRU(MORS)	0.80	**0.55**	0.85	**0.54**	**0.75**	**0.60**	**0.73**	**0.62**
LSTM	0.71	0.72	0.75	0.67	0.90	0.65	0.79	1.10
LSTM(MORS)	0.81	**0.52**	0.86	**0.54**	**0.76**	**0.57**	**0.71**	**0.60**

5.2.3.3 Custom Word2Vec Pre-Trained on the dataset Word Embeddings

Model	SC	CSS	FB	GS	IFE	VFM	Wi-Fi	UR
CNN	0.76	0.80	0.82	0.65	0.92	0.71	0.76	1.27
CNN(MORS)	**0.73**	**0.56**	**0.79**	**0.57**	**0.68**	**0.59**	**0.65**	**0.62**
GRU	0.74	0.74	0.77	0.61	0.86	0.65	0.79	1.15
GRU(MORS)	**0.70**	**0.59**	**0.75**	**0.58**	**0.68**	**0.63**	**0.65**	**0.64**
LSTM	0.72	0.73	0.74	0.65	0.86	0.65	0.77	1.09
LSTM(MORS)	0.74	**0.55**	0.83	**0.55**	**0.71**	**0.59**	**0.66**	**0.61**

In the case of deep learning models, since the word embeddings are all pre-trained, each of the models learns the same feature representations 7. Contrary to the inferior performance of TF-IDF Vectorizer, which tends to emphasize general phrases that might not capture specific insights of each category, MORS' with all pre-trained word embeddings expansive vocabulary is impressive.

This difference highlights MORS' capacity to produce a variety of expressions for every aspect(more useful representations) and the ability to extract precise information for each category, which increases the precision of the deep learning models compared to the machine learning models. Furthermore, the coefficient/feature importance analysis highlights that the keywords are dependent on how many reviews are considered from the dataset 4.2, since in all three word embedding techniques, the embeddings are pre-trained and this can be verified when considering "Value For Money" with "User Rating" and "Wi-Fi & Connectivity" with MORS, in which they have some common feature keywords because of the almost identical number of nullable reviews respectively. This implies that the above case(number of considered reviews) affects the obtained results(MAE) to some extent.

However, undoubtedly across various categories, the utilization of MORS consistently yields a notable decrease in prediction error when contrasted with the single-output models in almost the same categories for all word embedding techniques, except for some special cases e.g the "Seat Comfort" category in the Custom Word Embeddings set of experiments. For instance, it is pretty clear from 4.1 that even though the "Wi-Fi & Connectivity" category considers approximately the same number of reviews with MORS, superior efficiency in discerning nuanced patterns of passengers' sentiments is observed. When utilizing MORS "Wi-Fi & Connectivity" prediction error is decreased dramatically on every experiment(models & word embeddings). The interpretation superiority is also observed when it comes to the more complex "User Rating" category(10 labels considered), where MORS outperforms significantly every Single-Output Regression case.

5.2.4 Transformers

5.2.4.1 DistilBERT Tokenizer

Model	SC	CSS	FB	GS	IFE	VFM	Wi-Fi	UR
DistilBERT	0.72	0.67	0.75	0.60	0.81	0.62	0.71	1.27
DistilBERT(MORPS)	**0.70**	**0.57**	**0.72**	**0.56**	**0.66**	**0.60**	**0.61**	**0.61**

Contrary to the above set of experiments, the pre-trained DistilBERT tokenizer is not affected by the number of considered reviews, since every category in both Single/Multi-Output regression reveals the same feature representation keywords. When it comes to the rest of the models transitioning from single-output to MORS, the incorporation of multiple aspects might lead the models to recalibrate their representation of words to cater to the nuances of each category. On the other hand, unlike Word2Vec or GloVe, which you can extract and use as standalone vectors, DistilBERT learns contextualized representations of word embeddings as part of its pre-training process. The model's attention mechanism takes the learned word embeddings as input and uses them to create contextualized embeddings for each word based on where it appears in the sentence and how it interacts with other words.

In essence, the deep learning models recalibrate their understanding of words for distinct categories, while DistilBERT's self-attention mechanism inherently maintains consistent insights across regression tasks, enhancing its capacity to capture holistic contextual patterns which explains the improved performance. Lastly, the interpretability of MORS with DistilBERT plays a crucial role, demonstrating the model's potential for understanding complex feedback patterns and improving prediction precision as it outperforms every score of the Single-Output Regression cases. With MORS, the model is guided to simultaneously predict multiple aspects of passenger feedback, which encourages it to learn distinct linguistic patterns associated with each category.

5.2.5 Feature's Coefficients

This section discusses the obtained keywords from the coefficient/feature importance analysis in a meaningful manner and analyzes the extracted feature names by giving some insights. For the analytical observation of all the experiment results that coefficient/feature importance analysis produced visit the 7.

5.2.5.1 TF-IDF Vectorizer

Passengers' opinions on their travel experiences encompass a spectrum of emotions and assessments since as mentioned previously TF-IDF Vectorizer demonstrated more generic representations of the customer's feelings rather than aspect-specific features for each category. When considering "Seat Comfort," it's evident that passengers have encountered a range of situations from "uncomfortable" and "cramped" to those who found their seats "comfortable" and "grateful" for the space. Similarly, "Cabin Staff Service" appears to have had mixed impressions, with some finding the service "rude" and "worst," while others encountered staff who were "indifferent" or even "disinterested". "Food & Beverages" also had its disparities, with passengers experiencing both "inedible" and "poor" options, while others relished meals they found "excellent" and "delicious." Moving on to "Ground Service," passengers' experiences ranged from expressing "thanks" and feeling "grateful" for the help received, while others weren't as satisfied and conveyed their concerns. The "Inflight Entertainment" category showcased diversity, where some found the entertainment options "limited," while others appreciated the "excellent" choices offered. Assessing "Value For Money," opinions ranged from those who praised the value as "great" to those who considered it a "monopoly" with room for improvement. For "Wifi-Connectivity," opinions skewed positively, with passengers expressing their delight at the "excellent," "amazing," and "superb" connectivity. In the realm of "Overall User Rating," passengers offered a mix of "thanks," "great" and "excellent" ratings, while others experienced less satisfactory encounters. Finally, the MORS category reflected passengers' gratitude and praise for the "excellent" and "professional" service received, but there were those who encountered experiences they considered "worst." It seems that MORS considered the most important features of the above categories and mixed them in a way that a complete flight experience can be represented/described.

5.2.5.2 GloVe Pre-Trained Word Embeddings

In the context of GloVe pre-trained word embeddings, it's critical to understand that the models' choice of embeddings is influenced by the size of the training dataset. This affects the representation of feature names. For instance, in the "Seat Comfort" category, words like "striking" and "horrendously" appear, potentially representing encounters with seats. In "Cabin Staff Service", words like 'man', 'striking', and 'grumpy' emerge, providing insights into passenger-crew interactions. In "Food & Beverages", keywords like 'worthwhile,' 'juices,' and 'eats' are used to describe perceptions of meals. Terms like "annoy", "success", and "stopped" in "Ground Service" reflect attitudes toward airport services. "Inflight entertaining" uses words like "hazardous", "pressure", and "offload", all of which refer to entertaining activities. "Value For Money" links to places like "Dubrovnik" and "Moscow," giving away the traveler"s choices. "Wi-Fi & Connectivity" uses words like "twisted", "ransom", and "weak", which reflect Wi-Fi viewpoints. Last but not least, "Overall User Rating" summarizes overall opinions using phrases like "attention" "struggles", and "judgment". Notably, when utilizing MORS emphasizes words like "astronomical", "Philadelphia", "generation", and "Mercedes", suggesting attitudes influenced by the sample size of the training data.

5.2.5.3 FastText Pre-Trained Word Embeddings

The same point from above holds on FastText embeddings as well. When analyzing "Seat Comfort," terms like "strapped" and "immaculate" provide insights into seating quality and comfort perceptions. These keywords highlight ergonomic considerations and hygiene"s impact on overall comfort. In the "Cabin Staff Service" category, terms like "promote," "immaculate," "velocity," and "livery" mirror judgments about cabin crew conduct and presentation. They reveal how interactions and aesthetics contribute to staff service impressions. In "Food & Beverages," terms like "awfully," "bell," and "proposition" encapsulate sentiments about onboard meal quality and variety, emphasizing culinary experiences. "Ground Service" terms like "refunding" and "rerouting" reflect passenger interactions during emergencies, offering insights into problem-solving effectiveness. "In-Flight Entertainment" includes terms like "server" and "toasted," capturing the transformative role of entertainment. "Value For Money" gains dimension through words like "considering" and "sights," signifying value evaluations. "Wifi & Connectivity" terms like "luxurious" and "registration" mirror opinions on in-flight connectivity. "Overall User Rating" words like "dimensions" and "struggles" encapsulate overall sentiments. MORS-extracted terms like "airline" and "compensation" offer insights into broader feedback aspects, emphasizing a comprehensive understanding of passenger sentiments.

5.2.5.4 Custom Dataset Word Embeddings

In the "Seat Comfort" category, terms like "helped," "love," and "pacific" reflect passengers" seat comfort sensations. Similarly, "treat" and "cake" in "Cabin Staff Service" highlight interactions with the crew. In "Food & Beverages," words like "refund," "comforts," "discount," and "generous" convey opinions on meal quality and offerings, including discounts. "Ground Service" terms like "items" and "shuttle" relate to airport services. "Inflight Entertainment" terms like "blankets" and "arrangements" mirror opinions on entertainment. "Value For Money" connects to "details," "hostess," "enjoyed," and "improve," reflecting passenger treatment. "Wifi & Connectivity" terms like "safe," "satisfying," and "steal" depict Wi-Fi perceptions. "Overall User Rating" includes words like "purchase," "rate," and "upgraded," indicating general assessments. MORS-generated phrases like "peak," "status," and "disrespectfully" offer nuanced insights into passengers" experiences. MORS"s word embedding techniques reduce errors, demonstrating its ability to capture sentiment patterns and improve experience predictions.

5.2.5.5 DistilBERT Tokenizer

The list of prominent features extracted from the feature importance analysis of the DistilBERT model presents a captivating glimpse into the multifarious dimensions of passenger experiences. While the specific weight values are omitted, the identified terms like "elegance," "precautions" and "notify" imply a granularity in sentiment discernment. The incorporation of these distinct terms indicates the model's proficiency in capturing idiosyncratic aspects of passenger feedback. Moreover, the presence of "effortlessly," "ravaged," and "hierarchical" underscores the model's competence in elucidating an intricate spectrum of passenger encounters that profoundly influence their evaluative judgments. This assemblage of features not only underscores the model's prowess in unearthing nuanced linguistic nuances but also accentuates the strategic relevance for airlines to tailor their services in response to these intrinsic aspects. In this context, the top features emerge as a mosaic of semantic clues that can potentially inform strategic service enhancements based on the intricate tapestry of passenger perceptions.

Chapter 6

Conclusion

In the antecedent section, an exposition of the experimental evaluation of MORS has been delineated, where its performance has been juxtaposed with a spectrum of models that prognosticate singular ratings. In this ensuing chapter, a comprehensive synthesis of the thesis culminates through an exploration of the conclusions derived from the experimental evaluation. This is succeeded by a discerning analysis of the constraints encountered during the study, followed by an elucidation of the potential trajectories for future work emanating from the ambit of this project.

6.1 Summary & Conclusions

According to this work, adopting a **Multi-Output Regression System (MORS)** to combine all pertinent ratings for airlines' various services, enhances predictive performance. MORS generates more precise forecasts for particular features and, crucially, for the overall user rating when compared to single-output models. The overall user rating is influenced by the scores for several factors, including seat comfort, cabin staff service, etc, which seems plausible given the fact that all rating categories have high Pearson similarity between them 4.1. MORS successfully identifies these connections and raises the user rating's overall predictive accuracy. The interaction between various components is documented through MORS, emphasizing the dependencies and connections between them making it possible to capture nuanced insights specific to each aspect. For example, there is a significant correlation between "Value For Money" and "Ground Service", and this correlation affects the overall customer rating. These observations highlight the need of addressing various aspects comprehensively when assessing customer experiences.

Accurately assessing the grading procedure in light of these variables is complicated. The subjective nature of customer feedback is one of the biggest obstacles. Personal preferences, cultural variations, and a variety of experiences can all contribute to this subjectivity. There is a significant issue in quantifying the huge range of emotions and experiences. Additionally, the method of feature extraction and data preparation might not completely capture the nuanced feelings of passengers. Even though the study uses an advanced modeling methodology, it's important to recognize that some subtleties of consumer feedback might not be adequately captured by the attributes employed for prediction. As a result, there will be some approximation in the grading process. Furthermore, the challenge of interrelated aspects arises. Passengers' ratings for various aspects are often interconnected; for instance, their perception of cabin staff service

might impact their overall user rating. Capturing these intricate relationships and ensuring they are correctly represented in the grading process is challenging. Despite these challenges, the study's approach of utilizing Multi-Output Regression Systems (MORS) aids in addressing these issues to a pretty good extent. By considering multiple aspect scores simultaneously, MORS can capture the interdependencies and correlations among different aspects. However, the challenge of accurately interpreting and weighing these complex interrelations on this dataset persists.

The study uses a calculated strategy for dealing with incomplete reviews and missing data. A more thorough data cleaning procedure is used when switching from single-output to MORS. By deleting data instances with null values across all score columns, this method seeks to create a homogeneous training environment for multi-output models. Although this leads to more aggressive data elimination, it keeps the balance between data amount and suitability for multi-output learning. Depending on the method utilized, missing data might affect forecasts in a variety of ways. The comprehensive modeling of many different variables in MORS helps to lessen the impact of missing data for specific components. The algorithm can nevertheless make precise predictions for the overall user rating even in the absence of individual aspect scores by taking into account many scores at once. This demonstrates the advantage of MORS in handling incomplete data more effectively and maintaining robust prediction capabilities.

In conclusion, this study extensively investigates the factors influencing customer ratings of different aspects of airline services. The adoption of Multi-Output Regression Systems (MORS) significantly enhances the prediction accuracy of aspect ratings and overall user ratings. By combining relevant scores for various aspects, the relationships between these scores are effectively captured, providing a more holistic representation of passenger experiences. Furthermore, the study addresses challenges related to missing data and data preprocessing, emphasizing the importance of both accurate feature representation and interpretability. The insights gained from feature importance analysis shed light on passengers' sentiments and perceptions across different aspects, contributing to a comprehensive understanding of airline reviews. Through meticulous experimentation and analysis, this research enhances our understanding of passenger sentiments, aspect interdependencies, and the overall user rating determination process. The findings underscore the potential for more accurate and nuanced sentiment prediction, ultimately benefiting the airline industry by enabling targeted service enhancements and improved customer experiences.

6.2 Limitations

The master's thesis study was subject to some limitations that could have affected the accuracy and dependability of the target score prediction models. The existence of a large number of null scores in various dataset columns posed a substantial problem and had a considerable impact on the data preparation procedure. Careful consideration was needed to manage the missing factors and deal with the uneven distribution of scores. Interesting patterns emerged from the score distribution analysis, with some labels being evenly distributed across the scale and others being over-represented. Filling empty spaces with average values may have induced bias and inaccurately reflect the genuine sentiments of passengers toward particular issues. This strategy would make it more difficult for the models to faithfully represent user sentiment and differences in their experiences, which would lead to less accurate findings.

The partial scope of the experimental baseline models and word embedding techniques used were another drawback. The study may have missed chances for even better outcomes and a more in-depth understanding of the data by not examining a wider range of cutting-edge models and

more complex embeddings. Additionally, the exclusive use of particular word embedding schemes may have disregarded other techniques that might have represented the semantics and context of the reviews more accurately. Integrating sophisticated transformer-based architectures, such as BERT, ELECTRA, and RoBERTa, might have produced representations of the text that are more contextually aware and enhanced performance on tasks requiring natural language understanding, given the fact that the transformer's performance in this work demonstrated great results. However, the time needed for training such complex/large models and investigating additional word embedding techniques restricted the investigation of more complicated transformer topologies. Despite their potential advantages, a lack of resources prohibited a thorough analysis of their impact on the study's goals. Incorporating more sophisticated transformer models could have provided more information about passenger perceptions and improved the models' forecasting ability.

6.3 Future Work

To minimize prejudice and preserve the objectivity of passenger opinions, it is crucial to research various methods for handling missing values. The manual labeling of empty score cells is a useful but time-consuming strategy for the stability of the dataset. Techniques like expert-judged manual imputation or potent machine learning models that account for missingness could be taken into consideration. Any imputation strategy must be carefully evaluated and validated to make sure it does not generate unanticipated biases or distortions in the data. The research's credibility is increased and more meaningful information regarding passenger experiences is provided by openly disclosing the processes used to handle missing values in the study. This transparency also enables readers to comprehend the potential influence on findings and interpretations. Additionally, future studies must look into a larger range of baseline models, word embedding techniques, and transformer designs. Results that are more dependable and exact can be obtained by combining a variety of approaches with the most recent NLP advancements.

In the context of this work, model stacking has a great deal of potential as a promising area for future research. Model stacking, sometimes referred to as ensemble learning, mixes various prediction models to improve overall performance and predictive accuracy. The limitations discovered in this study are effectively addressed by stacking by utilizing the benefits of numerous models. By combining different baseline models and experimenting with various word embedding methodologies, it can precisely capture even more complicated patterns and nuanced parts of passenger reviews. Model stacking can also reduce the consequences of missing data and imbalanced score distributions by combining the advantages of various models.

The use of advanced transformer-based embeddings in model stacking, such as BERT, ELECTRA, and RoBERTa, may produce representations of textual data that are more thorough and contextually aware. Model stacking can create more accurate and perceptive analyses of passenger sentiment and experiences in the aviation sector by enhancing the generalization and resilience capabilities of prediction models. Model stacking thus presents a promising direction for future study to deepen the comprehension of passenger feelings and increase the forecasting capability of the models.

Bibliography

Alvarez, M., & Lawrence, N. (2008). Sparse convolved gaussian processes for multi-output regression. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 21). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2008/file/149e9677a5989fd342ae44213df68868-Paper.pdf

Bishop, C. M. (2007). *Pattern recognition and machine learning (information science and statistics)* (1st ed.). Springer. Retrieved from http://www.amazon.com/Pattern-Recognition-Learning-Information-Statistics/dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. Retrieved from https://aclanthology.org/Q17-1010 doi: 10.1162/tacl_a_00051

Borchani, H., Varando, G., Bielza, C., & Larranaga, P. (2015). A survey on multi-output regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *5*(5), 216–233.

Charatsaris, G. (2017). *Efficient aspect based sentiment analysis with application to airline reviews.*

Chen, C., Zhuo, R., & Ren, J. (2019, 06). Gated recurrent neural network with sentimental relations for sentiment classification. *Information Sciences*, *502*. doi: 10.1016/j.ins.2019.06.050

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1179 doi: 10.3115/v1/D14-1179

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014, 12). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Dereli, N., & Saraclar, M. (2019, July). Convolutional neural networks for financial text regression. In *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop* (pp. 331–337). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-2046 doi: 10.18653/v1/P19-2046

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, 10). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Eslami, P., Ghasemaghaei, M., & Hassanein, K. (2018, 07). Which online reviews' do consumers find most helpful?: A multimethod investigation. *Decision Support Systems*, *113*. doi: 10.1016/j.dss.2018.06.012

Fang, B., Ye, Q., Kucukusta, D., & Law, R. (2016, 02). Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management*, *52*, 498-506. doi: 10.1016/j.tourman.2015.07.018

Giannakopoulos, A., Musat, C., Hossmann, A., & Baeriswyl, M. (2017, September). Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets. In *Proceedings of the 8th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 180–188). Copenhagen, Denmark: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W17-5224` doi: 10.18653/v1/W17-5224

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY, USA: Springer New York Inc.

He, R., Lee, W. S., Ng, H. T., & Dahlmeier, D. (2019, July). An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 504–515). Florence, Italy: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P19-1048` doi: 10.18653/v1/P19-1048

Hochreiter, S., & Schmidhuber, J. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. doi: 10.1162/neco.1997.9.8.1735

Jacovi, A., Sar Shalom, O., & Goldberg, Y. (2018, November). Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 56–65). Brussels, Belgium: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W18-5408` doi: 10.18653/v1/W18-5408

Joshy, A., & Sundar, S. (2022). Analyzing the performance of sentiment analysis using bert, distilbert, and roberta. In *2022 ieee international power and renewable energy conference (iprecon)* (p. 1-6). doi: 10.1109/IPRECON55716.2022.10059542

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014, June). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 655–665). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P14-1062` doi: 10.3115/v1/P14-1062

Li, R., Lin, Z., Lin, H., Wang, W., & Meng, D. (2018, 01). Text emotion analysis: A survey. *Jisuanji Yanjiu yu Fazhan/Computer Research and Development*, *55*, 30-52. doi: 10.7544/issn1000-1239.2018.20170055

Liu, Z., & Park, S. (2015). What makes a useful online review? implication for travel product websites. *Tourism Management*, *47*, 140-151. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0261517714001903` doi: https://doi.org/10.1016/j.tourman.2014.09.020

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Mouselimis, L. (2022). fastText: Efficient learning of word representations and sentence classification using r [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=fastText` (R package version 1.0.3)

Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Park, S., Lee, J., & Nicolau, J. (2020, 12). Understanding the dynamics of the quality of airline service attributes: Satisfiers and dissatisfiers. *Tourism Management*, *81*, 104163. doi: 10.1016/j.tourman.2020.104163

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D14-1162 doi: 10.3115/v1/D14-1162

Sachin, S., Tripathi, A., Mahajan, N., Aggarwal, S., & Nagrath, P. (2020, 03). Sentiment analysis using gated recurrent neural networks. *SN Computer Science*, *1*. doi: 10.1007/s42979-020-0076-y

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Santur, Y. (2019, 09). Sentiment analysis based on gated recurrent unit. In (p. 1-5). doi: 10.1109/IDAP.2019.8875985

Sezgen, E., Mason, K., & Mayer, R. (2019, 06). Voice of airline passenger: A text mining approach to understand customer satisfaction. *Journal of Air Transport Management*, *77*, 65-74. doi: 10.1016/j.jairtraman.2019.04.001

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd international conference on world wide web* (p. 373–374). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2567948.2577348 doi: 10.1145/2567948.2577348

Song, C., Guo, J., & Zhuang, J. (2020, 10). Analyzing passengers' emotions following flight delays- a 2011–2019 case study on skytrax comments. *Journal of Air Transport Management*, *89*, 101903. doi: 10.1016/j.jairtraman.2020.101903

Support Vector Regressor svr. (n.d.-a). https://scikit-learn.sourceforge.net/0.5/auto_examples/svm/plot_svm_regression.html. (Accessed: 2010-09-30)

Support Vector Regressor svr. (n.d.-b). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html. (Accessed: 2010-09-30)

Talaat, A. S. (2023). Sentiment analysis classification system using hybrid bert models. *Journal of Big Data*, *10*(1), 1–18.

Tang, D., Qin, B., Feng, X., & Liu, T. (2016, December). Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 3298–3307). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://aclanthology.org/C16-1311

Tang, D., Qin, B., & Liu, T. (2016, November). Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 214–224). Austin, Texas: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D16-1021 doi: 10.18653/v1/D16-1021

Tf–idf. (2010). In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of machine learning* (pp. 986–987). Boston, MA: Springer US. Retrieved from https://doi.org/10.1007/978-0-387-30164-8_832 doi: 10.1007/978-0-387-30164-8_832

Tiwari, P., Yadav, P., Kumar, S., Mishra, B., Nhu, N., Gochhayat, S., ... Prasad, M. (2018, 11). Sentiment analysis for airlines services based on twitter dataset. In (p. 14). doi: 10.1016/B978-0-12-815458-8.00008-6

Verma, K., & Davis, B. (2021, 07). Implicit aspect-based opinion mining and analysis of airline industry based on user-generated reviews. *SN Computer Science*, *2*. doi: 10.1007/s42979-021-00669-7

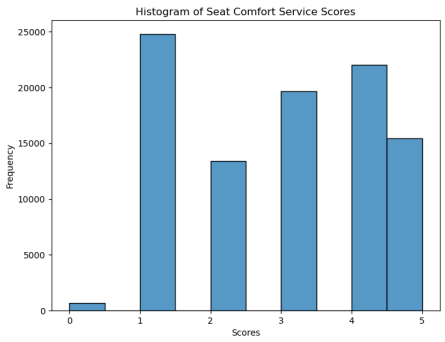
Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016, November). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 606–615). Austin, Texas: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D16-1058` doi: 10.18653/v1/D16-1058

Yih, W.-t., He, X., & Meek, C. (2014, June). Semantic parsing for single-relation question answering. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 643–648). Baltimore, Maryland: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P14-2105` doi: 10.3115/v1/P14-2105

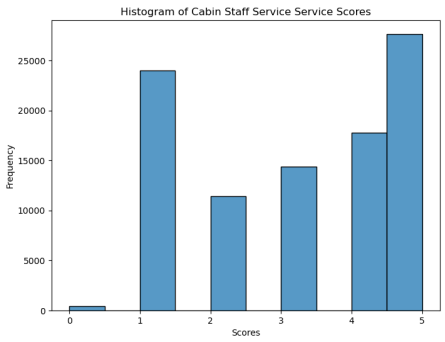
Chapter 7

Appendix

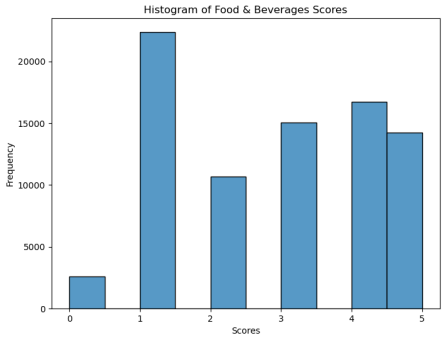
7.1 Histograms of scores without MORS



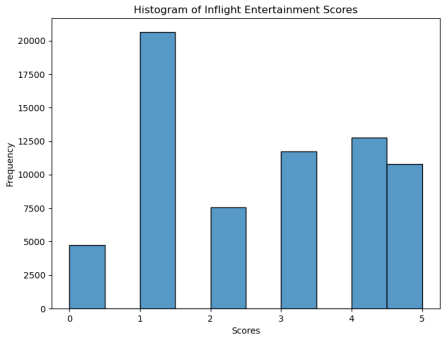
(a) Seat Comfort Scores Histogram.



(b) Cabin Staff Service Scores Histogram.

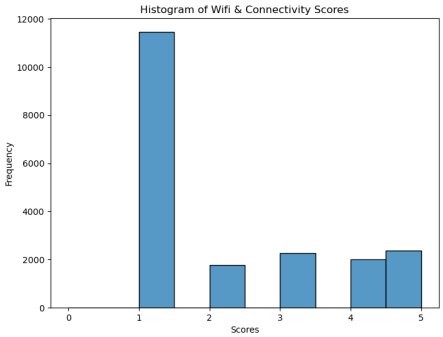


(c) Food & Beverages Scores Histogram.

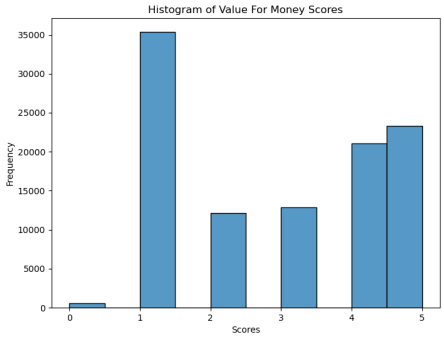


(d) Inflight Entertainment Scores Histogram.

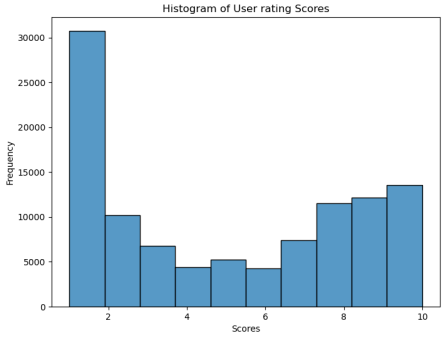
Figure 7.1: Page 1 of 2 with scores distrubution histograms.



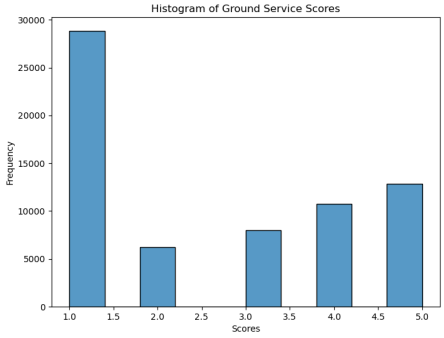
(e) Wifi & Connectivity Scores Histogram.



(f) Value For Money Scores Histogram.



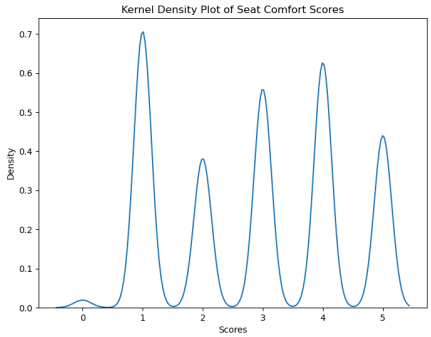
(g) User Rating Scores Histogram.



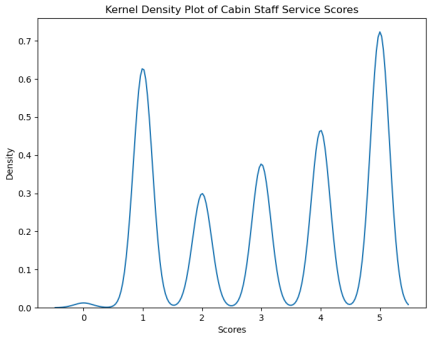
(h) Ground Service Scores Histogram.

Figure 7.1: Page 2 of 2 with scores distrubution histograms.

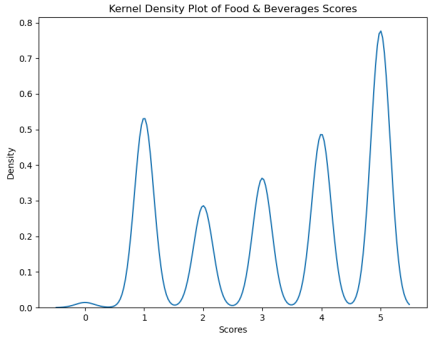
7.2 Kernel Density Functions of scores without MORS



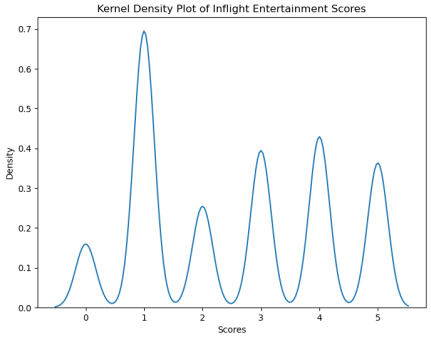
(a) Seat Comfort Scores Kernel Density Function.



(b) Cabin Staff Service Scores Kernel Density Function.

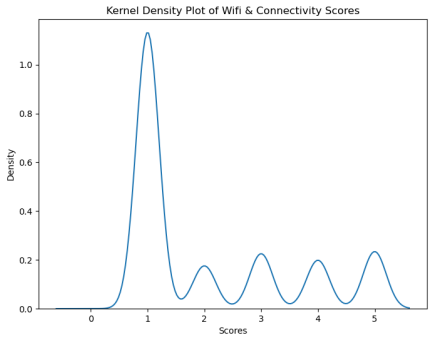


(c) Food & Beverages Scores Kernel Density Function.

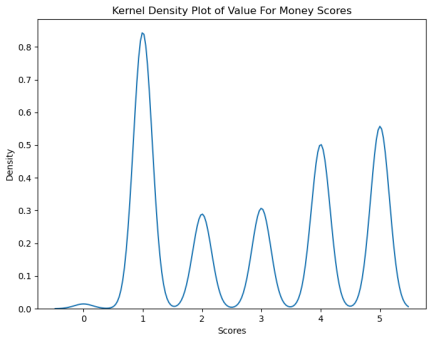


(d) Inflight Entertainment Scores Kernel Density Function.

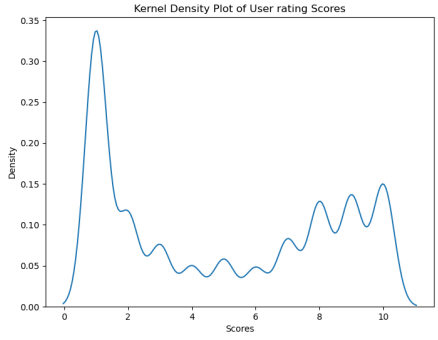
Figure 7.2: Page 1 of 2 with scores kernel density functions.



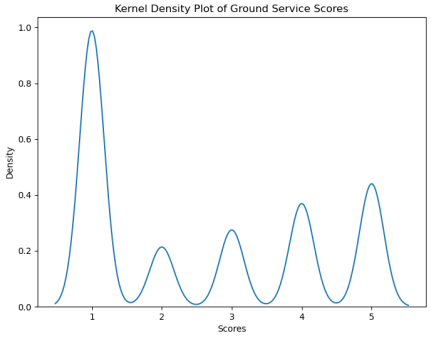
(e) Wifi & Connectivity Scores Kernel Density Function.



(f) Value For Money Scores Kernel Density Function.



(g) User Rating Scores Kernel Density Function.



(h) Ground Service Scores Kernel Density Function.

Figure 7.2: Page 2 of 2 with scores kernel density functions.

7.3 Descriptive statistics of scores without MORS

count	95966.000000
mean	2.873862
std	1.447085
min	0.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000

(a) Seat Comfort Scores descriptive statistics.

count	95800.000000
mean	3.127630
std	1.578235
min	0.000000
25%	1.000000
50%	3.000000
75%	5.000000
max	5.000000

(b) Cabin Staff Service Scores descriptive statistics.

count	81708.000000
mean	2.779373
std	1.538304
min	0.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000

(c) Food & Beverages Scores descriptive statistics.

count	68231.000000
mean	2.578842
std	1.609229
min	0.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000

(d) Inflight Entertainment Scores descriptive statistics.

Figure 7.3: Page 1 of 2 with scores descriptive statistics.

count	19890.000000
mean	2.096682
std	1.468408
min	0.000000
25%	1.000000
50%	1.000000
75%	3.000000
max	5.000000

(e) Wifi & Connectivity Scores descriptive statistics.

count	105264.000000
mean	2.838986
std	1.603252
min	0.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000

(f) Value For Money Scores descriptive statistics.

count	17988.000000
mean	3.919280
std	3.398553
min	1.000000
25%	1.000000
50%	2.000000
75%	7.000000
max	10.000000

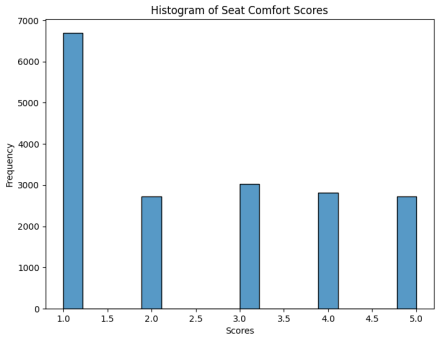
(g) User Rating Scores descriptive statistics.

count	66717.000000
mean	2.588531
std	1.607866
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

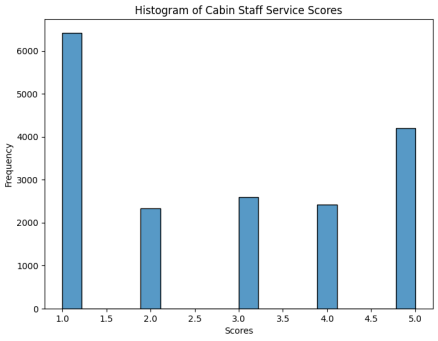
(h) Ground Service Scores descriptive statistics.

Figure 7.3: Page 2 of 2 with scores descriptive statistics.

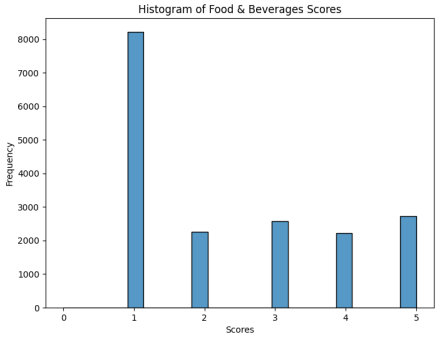
7.4 Histograms of scores with MORS



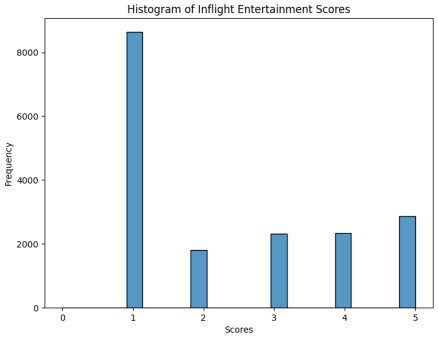
(a) Seat Comfort Scores Histogram (MORS).



(b) Cabin Staff Service Scores Histogram (MORS).

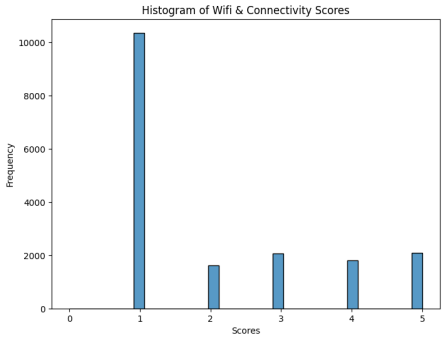


(c) Food & Beverages Scores Histogram (MORS).



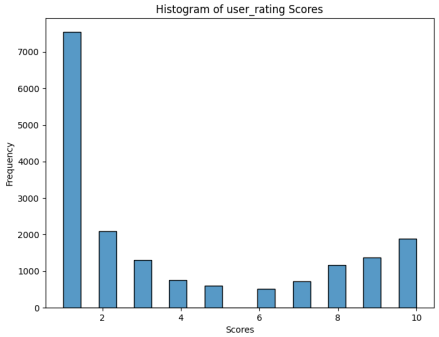
(d) Inflight Entertainment Scores Histogram (MORS).

Figure 7.4: Page 1 of 2 with scores distrubution histograms (MORS).

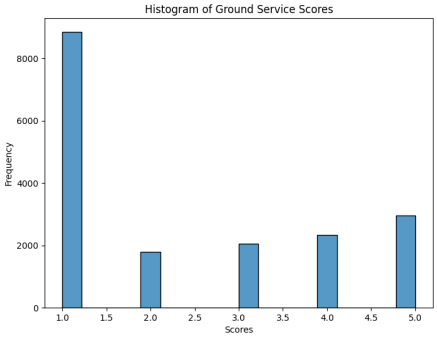


(e) Wifi & Connectivity Scores Histogram (MORS).

(f) Value For Money Scores Histogram (MORS).



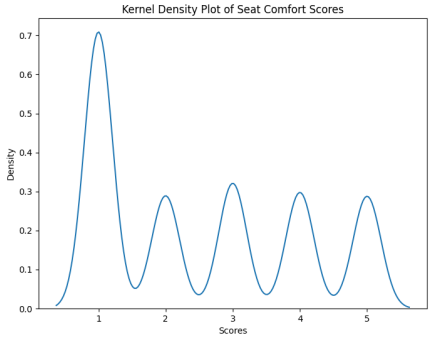
(g) User Rating Scores Histogram (MORS).



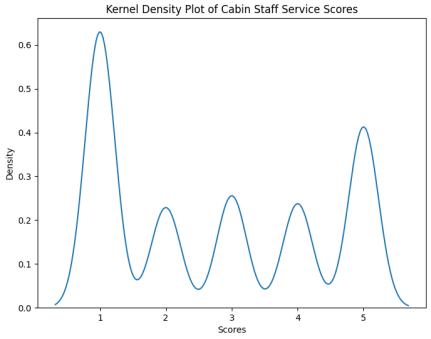
(h) Ground Service Scores Histogram (MORS).

Figure 7.4: Page 2 of 2 with scores distrubution histograms (MORS).

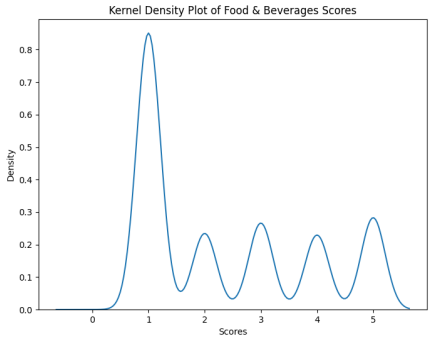
7.5 Kernel Density Functions of scores with MORS



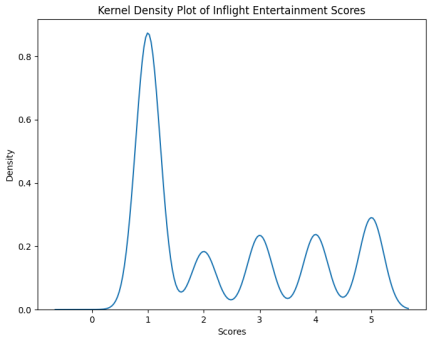
(a) Seat Comfort Scores Histogram (MORS).



(b) Cabin Staff Service Scores Histogram (MORS).

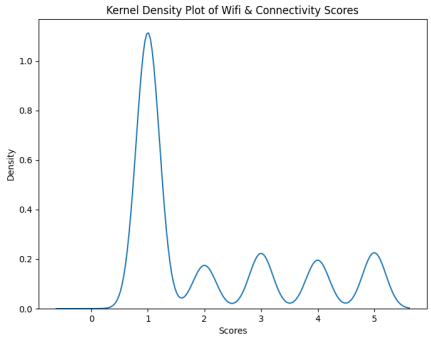


(c) Food & Beverages Scores Histogram (MORS).

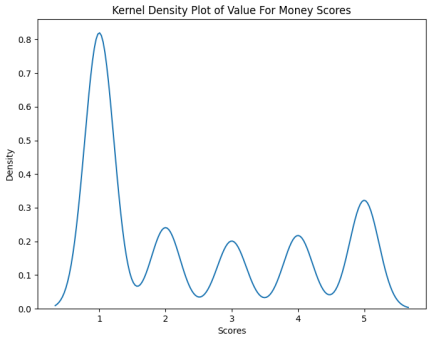


(d) Inflight Entertainment Scores Histogram (MORS).

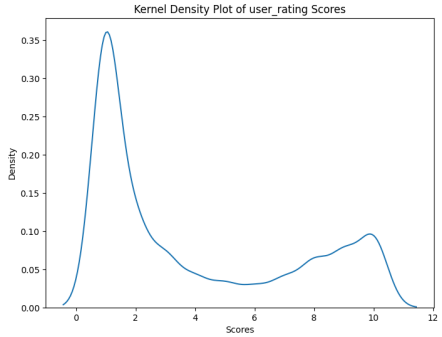
Figure 7.5: Page 1 of 2 with scores kernel density functions (MORS).



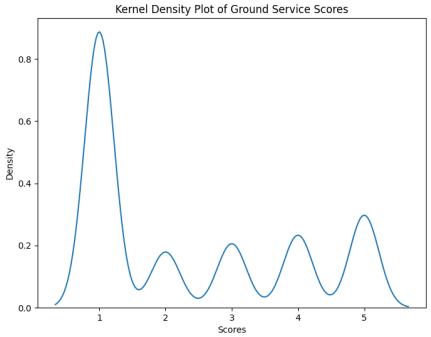
(e) Wifi & Connectivity Scores Histogram (MORS).



(f) Value For Money Scores Histogram (MORS).



(g) User Rating Scores Histogram (MORS).



(h) Ground Service Scores Histogram (MORS).

Figure 7.5: Page 2 of 2 with scores kernel density functions (MORS).

7.6 Descriptive Statistics of scores with MORS

count	17988.000000
mean	2.562097
std	1.486614
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

(a) Seat Comfort Scores Histogram (MORS).

count	17988.000000
mean	2.758950
std	1.603195
min	1.000000
25%	1.000000
50%	3.000000
75%	4.000000
max	5.000000

(b) Cabin Staff Service Scores Histogram (MORS).

count	17988.000000
mean	2.387203
std	1.519030
min	0.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

(c) Food & Beverages Scores Histogram (MORS).

count	17988.000000
mean	2.388092
std	1.554762
min	0.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

(d) Inflight Entertainment Scores Histogram (MORS).

Figure 7.6: Page 1 of 2 with scores descriptive statistics with (MORS).

count	17988.000000
mean	2.091617
std	1.462700
min	0.000000
25%	1.000000
50%	1.000000
75%	3.000000
max	5.000000

(e) Wifi & Connectivity Scores Histogram (MORS).

count	17988.000000
mean	2.434067
std	1.571746
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

(f) Value For Money Scores Histogram (MORS).

count	17988.000000
mean	3.919280
std	3.398553
min	1.000000
25%	1.000000
50%	2.000000
75%	7.000000
max	10.000000

(g) User Rating Scores Histogram (MORS).

count	17988.000000
mean	2.375473
std	1.570954
min	1.000000
25%	1.000000
50%	2.000000
75%	4.000000
max	5.000000

(h) Ground Service Scores Histogram (MORS).

Figure 7.6: Page 2 of 2 with scores descriptive statistics with (MORS).

7.7 Support Vector Regression Coefficient Analysis

	Feature	Coefficient	AbsoluteCoefficient
59519	uncomfortable	-6.031707	6.031707
13317	cramped	-4.183525	4.183525
38277	narrow	-3.831030	3.831030
63267	worst	-3.821243	3.821243
11654	comfortable	3.613888	3.613888
39486	not	-3.112243	3.112243
38664	never	-2.803845	2.803845
24126	grateful	2.664896	2.664896
56849	thank	2.644206	2.644206
57326	tight	-2.630644	2.630644

(a) Seat Comfort Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
48885	rude	-4.853324	4.853324
63160	worst	-4.501833	4.501833
59767	unfriendly	-4.126858	4.126858
39435	not	-3.817871	3.817871
27842	indifferent	-3.768191	3.768191
55225	surly	-3.640907	3.640907
16116	disinterested	-3.540841	3.540841
38612	never	-3.505675	3.505675
48531	robotic	-3.279359	3.279359
16814	downhill	-3.132656	3.132656
	Feature	Importance	

(b) Cabin Staff Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
26222	inedible	-4.818721	4.818721
59174	worst	-3.784871	3.784871
18365	excellent	3.665711	3.665711
37038	not	-3.608419	3.608419
13921	delicious	3.498594	3.498594
41270	poor	-3.377606	3.377606
52782	tasteless	-3.299957	3.299957
36575	no	-3.037050	3.037050
4060	awful	-2.948040	2.948040
52793	tasty	2.924395	2.924395

(c) Food & Beverages Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
43867	thank	4.889287	4.889287
20021	helped	3.761170	3.761170
18894	grateful	3.421853	3.421853
33766	praise	3.327737	3.327737
29882	not	-3.067430	3.067430
24114	kudos	3.049364	3.049364
38244	satisfied	2.918638	2.918638
15348	excellent	2.903558	2.903558
43869	thankful	2.808411	2.808411
49132	worried	2.805414	2.805414

(d) Ground Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
3218	atr	-4.222154	4.222154
28371	limited	-3.508849	3.508849
42643	scoot	-3.352769	3.352769
31868	monarch	-3.334988	3.334988
20593	great	3.315432	3.315432
23850	indigo	-3.241514	3.241514
15790	embraer	-3.184410	3.184410
16789	excellent	3.135389	3.135389
33404	no	-3.124915	3.124915
27358	lcct	-2.944565	2.944565

(e) Inflight Entertainment Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
60364	thank	3.535697	3.535697
42179	not	-3.509297	3.509297
67068	worst	-3.164865	3.164865
41278	never	-3.080183	3.080183
25635	grateful	3.080087	3.080087
20935	excellent	2.908125	2.908125
2194	amazing	2.606859	2.606859
47295	praise	2.575861	2.575861
39796	monopoly	-2.568297	2.568297
25672	great	2.565829	2.565829

(f) Value For Money Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
8565	excellent	4.368639	4.368639
24462	thank	4.040356	4.040356
974	amazing	3.370648	3.370648
10565	great	3.370228	3.370228
17132	outstanding	2.964830	2.964830
13375	kudos	2.781276	2.781276
23760	superb	2.715346	2.715346
4764	comfortable	2.691179	2.691179
11945	impressed	2.641365	2.641365
17824	perfect	2.626001	2.626001

(g) Wifi & Connectivity Absolute Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
60605	thank	8.910926	8.910926
42320	not	-8.091000	8.091000
25667	grateful	7.596482	7.596482
67366	worst	-6.914140	6.914140
20924	excellent	6.879032	6.879032
25702	great	5.988942	5.988942
41434	never	-5.706643	5.706643
2197	amazing	5.655932	5.655932
63425	uncomfortable	-5.457702	5.457702
47127	poor	-5.362008	5.362008

(h) User Rating Coefficient Analysis.

Figure 7.7: Page of SVR coefficient analysis with top factors without MORS.

7.8 Ridge Regression Coefficient Analysis

	Feature	Coefficient	AbsoluteCoefficient
59519	uncomfortable	-6.092555	6.092555
13317	cramped	-4.062126	4.062126
38277	narrow	-3.672518	3.672518
63267	worst	-3.477484	3.477484
11654	comfortable	3.448672	3.448672
39486	not	-2.900248	2.900248
57326	tight	-2.706128	2.706128
8202	bundaberg	-2.682508	2.682508
19657	excellent	2.588965	2.588965
56849	thank	2.552818	2.552818

(a) Seat Comfort Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
48885	rude	-4.444039	4.444039
63160	worst	-3.869792	3.869792
59767	unfriendly	-3.797916	3.797916
16116	disinterested	-3.486780	3.486780
39435	not	-3.452372	3.452372
27842	indifferent	-3.408904	3.408904
19570	excellent	3.230908	3.230908
8155	bundaberg	-3.224694	3.224694
55225	surly	-3.187057	3.187057
27580	inattentive	-2.913829	2.913829

(b) Cabin Staff Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
26222	inedible	-5.059441	5.059441
59174	worst	-3.932038	3.932038
52782	tasteless	-3.613989	3.613989
18365	excellent	3.612369	3.612369
13921	delicious	3.544370	3.544370
52793	tasty	3.197666	3.197666
37038	not	-3.169699	3.169699
41270	poor	-3.115321	3.115321
7689	bundaberg	-2.803861	2.803861
4060	awful	-2.797867	2.797867

(c) Food & Beverages Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
43867	thank	3.970054	3.970054
18894	grateful	3.311427	3.311427
29882	not	-3.256718	3.256718
49148	worst	-3.159966	3.159966
20021	helped	3.144229	3.144229
37724	rude	-2.921392	2.921392
43869	thankful	2.796631	2.796631
7475	chaotic	-2.654090	2.654090
24114	kudos	2.640511	2.640511
15348	excellent	2.608792	2.608792

(d) Ground Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
31868	monarch	-3.444527	3.444527
3218	atr	-3.434377	3.434377
42643	scoot	-3.264640	3.264640
2112	ancient	-3.208816	3.208816
28371	limited	-3.081787	3.081787
18480	flybe	-3.065386	3.065386
27358	lcct	-3.036872	3.036872
20593	great	2.994547	2.994547
15790	embraer	-2.982038	2.982038
37739	porter	-2.924361	2.924361

(e) Inflight Entertainment Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
67068	worst	-3.412659	3.412659
42179	not	-3.319122	3.319122
60364	thank	3.226949	3.226949
25635	grateful	2.995936	2.995936
41278	never	-2.987319	2.987319
8797	bundaberg	-2.853897	2.853897
20935	excellent	2.797622	2.797622
2194	amazing	2.638195	2.638195
39796	monopoly	-2.619954	2.619954
60366	thankful	2.606381	2.606381

(f) Value For Money Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
8565	excellent	3.316395	3.316395
24462	thank	3.130399	3.130399
27349	worst	-2.665140	2.665140
16417	not	-2.571819	2.571819
16208	no	-2.444128	2.444128
10565	great	2.442827	2.442827
974	amazing	2.388843	2.388843
13375	kudos	2.337585	2.337585
17132	outstanding	2.315755	2.315755
16057	never	-2.277911	2.277911

(g) Wifi & Connectivity Absolute Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
25667	grateful	8.635444	8.635444
60605	thank	8.279579	8.279579
42320	not	-8.155306	8.155306
67366	worst	-7.711587	7.711587
55260	silverjet	-7.202617	7.202617
20924	excellent	6.679001	6.679001
60607	thankful	5.756876	5.756876
25702	great	5.691050	5.691050
2197	amazing	5.679044	5.679044
60467	terrible	-5.373751	5.373751

(h) User Rating Coefficient Analysis.

Figure 7.8: Page of Ridge Regression coefficient analysis with top factors without MORS.

7.9 Multi Ouput Regression Ridge & SVR

	Feature	Coefficient	AbsoluteCoefficient
208158	thank	7.774585	7.774585
192967	excellent	7.711426	7.711426
200507	not	-6.996154	6.996154
194887	great	6.979099	6.979099
210883	worst	-6.722648	6.722648
185670	amazing	6.288274	6.288274
201186	outstanding	5.880564	5.880564
193330	fantastic	5.695771	5.695771
187122	best	5.617258	5.617258
202946	professional	5.523131	5.523131

(a) Multi Output SVR Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
208158	thank	7.774585	7.774585
192967	excellent	7.711426	7.711426
200507	not	-6.996154	6.996154
194887	great	6.979099	6.979099
210883	worst	-6.722648	6.722648
185670	amazing	6.288274	6.288274
201186	outstanding	5.880564	5.880564
193330	fantastic	5.695771	5.695771
187122	best	5.617258	5.617258
202946	professional	5.523131	5.523131

(b) Multi Output Regression Ridge Coefficient Analysis.

7.10 All Deep Learning Models Custom Word2Vec Word Embeddings Coefficient Analysis

	Feature	Coefficient	AbsoluteCoefficient
554	dhabi	-12.817374	12.817374
1166	love	11.659348	11.659348
842	pacific	-11.041563	11.041563
828	helped	-10.649260	10.649260
1169	directly	10.318373	10.318373
4771	booth	10.301453	10.301453
1221	paper	-9.917017	9.917017
6536	generated	-9.895596	9.895596
6036	olds	9.808111	9.808111
1632	fill	-9.800575	9.800575

(a) (Custom Word Embeddings)Seat Comfort Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
554	las	-12.817374	12.817374
1165	till	11.659348	11.659348
848	recent	-11.041563	11.041563
828	happen	-10.649260	10.649260
1167	love	10.318373	10.318373
4768	booth	10.301453	10.301453
1220	treat	-9.917017	9.917017
6529	generated	-9.895596	9.895596
6029	olds	9.808111	9.808111
1632	cake	-9.800575	9.800575

Top 10

(b) (Custom Word Embeddings)Cabin Staff Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
497	refund	-12.817374	12.817374
1113	packed	11.659348	11.659348
952	nobody	-11.041563	11.041563
849	stand	-10.649260	10.649260
1083	generous	10.318373	10.318373
4857	win	10.301453	10.301453
1193	feet	-9.917017	9.917017
7084	insulted	-9.895596	9.895596
6247	comforts	9.808111	9.808111
1641	discount	-9.800575	9.800575

(c) (Custom Word Embeddings)Food & Beverages Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
623	tasty	-12.817374	12.817374
1246	shuttle	11.659348	11.659348
729	believe	-11.041563	11.041563
804	notice	-10.649260	10.649260
1278	item	10.318373	10.318373
4525	smelly	10.301453	10.301453
1140	st	-9.917017	9.917017
5640	dmk	-9.895596	9.895596
5319	nagoya	9.808111	9.808111
1624	likely	-9.800575	9.800575

(d) (Custom Word Embeddings)Ground Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
475	seems	-12.817374	12.817374
1098	running	11.659348	11.659348
1010	supervisor	-11.041563	11.041563
877	blankets	-10.649260	10.649260
1061	mention	10.318373	10.318373
4683	buggy	10.301453	10.301453
1201	list	-9.917017	9.917017
7580	remembering	-9.895596	9.895596
6539	applicable	9.808111	9.808111
1667	arrangements	-9.800575	9.800575

(e) (Custom Word Embeddings)Inflight Entertainment Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
554	purchase	-12.817374	12.817374
1197	whether	11.659348	11.659348
871	perth	-11.041563	11.041563
826	enjoyed	-10.649260	10.649260
1178	improve	10.318373	10.318373
4901	sticking	10.301453	10.301453
1206	details	-9.917017	9.917017
6418	denying	-9.895596	9.895596
5356	survey	9.808111	9.808111
1642	hostess	-9.800575	9.800575

(f) (Custom Word Embeddings)Value For Money Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
627	believe	-12.817374	12.817374
1393	higher	11.659348	11.659348
774	rows	-11.041563	11.041563
790	takeoff	-10.649260	10.649260
1221	safe	10.318373	10.318373
3968	satisfying	10.301453	10.301453
1385	afternoon	-9.917017	9.917017
5957	escalate	-9.895596	9.895596
4528	steal	9.808111	9.808111
1649	reservations	-9.800575	9.800575

(g) (Custom Word Embeddings)Wifi & Connectivity Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
555	purchase	-12.817374	12.817374
1198	rate	11.659348	11.659348
877	upgraded	-11.041563	11.041563
826	considering	-10.649260	10.649260
1175	improve	10.318373	10.318373
4923	dragon	10.301453	10.301453
1205	gets	-9.917017	9.917017
6273	fro	-9.895596	9.895596
5371	survey	9.808111	9.808111
1640	hostess	-9.800575	9.800575

(h) (Custom Word Embeddings)User Rating Coefficient Analysis.

7.11 All Deep Learning Models GloVe pre-trained Word Embeddings Coefficient Analysis

	Feature	Coefficient	AbsoluteCoefficient
9439	excursions	25.035873	25.035873
9502	horrendously	24.836399	24.836399
4530	buggy	-24.619127	24.619127
5182	dammam	23.226828	23.226828
1067	outside	-23.201000	23.201000
9619	striking	22.625372	22.625372
2704	grumpy	-21.603767	21.603767
7961	belly	-21.533394	21.533394
8213	shocker	-21.097683	21.097683
1114	five	-20.661144	20.661144

(a) (GloVe Word Embeddings)Seat Comfort Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
9434	excursions	25.035873	25.035873
9498	horrendously	24.836399	24.836399
4525	buggy	-24.619127	24.619127
5179	dammam	23.226828	23.226828
1064	man	-23.201000	23.201000
9619	striking	22.625372	22.625372
2702	grumpy	-21.603767	21.603767
7956	belly	-21.533394	21.533394
8206	shocker	-21.097683	21.097683
1115	expectations	-20.661144	20.661144

(b) (GloVe Word Embeddings)Cabin Staff Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
9851	ao	24.836399	24.836399
5077	worthwhile	-24.619127	24.619127
4771	yul	23.226828	23.226828
1014	passport	-23.201000	23.201000
8944	staggering	22.625372	22.625372
2833	juices	-21.603767	21.603767
7481	cloths	-21.533394	21.533394
9695	propose	21.165909	21.165909
8229	eats	-21.097683	21.097683
1302	rebook	-20.661144	20.661144

(c) (GloVe Word Embeddings)Food & Beverages Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
8211	annoy	25.035873	25.035873
8261	bothers	24.836399	24.836399
3867	success	-24.619127	24.619127
7168	antananarivo	23.226828	23.226828
1260	stopped	-23.201000	23.201000
2628	babies	-21.603767	21.603767
8138	jos	-21.533394	21.533394
9871	bloemfontein	-21.266447	21.266447
9032	dozed	21.165909	21.165909
7420	sullen	-21.097683	21.097683

(d) (GloVe Word Embeddings)Ground Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
5322	dangerous	-24.619127	24.619127
4400	england	23.226828	23.226828
1015	salad	-23.201000	23.201000
2926	pressure	-21.603767	21.603767
7147	doubts	-21.533394	21.533394
1353	foot	-20.661144	20.661144
6064	offload	20.640966	20.640966
683	eventually	-20.564966	20.564966
1418	local	-20.329384	20.329384
6534	clubs	20.148067	20.148067

(e) (GloVe Word Embeddings)Inflight Entertainment Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
9602	lasagna	25.035873	25.035873
9843	humiliation	24.836399	24.836399
4449	speaks	-24.619127	24.619127
5270	dubrovnik	23.226828	23.226828
1054	moscow	-23.201000	23.201000
9994	confuse	22.625372	22.625372
2791	assignment	-21.603767	21.603767
8039	wornout	-21.533394	21.533394
9801	sights	21.165909	21.165909
7779	verde	-21.097683	21.097683

(f) (GloVe Word Embeddings)Value For Money Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
8722	hanger	25.035873	25.035873
7845	ransom	24.836399	24.836399
3737	weak	-24.619127	24.619127
8297	mercedes	-23.308704	23.308704
8672	twisted	23.226828	23.226828
1318	picked	-23.201000	23.201000
2291	organization	-21.603767	21.603767
6733	intimate	-21.533394	21.533394
7950	shameless	-21.097683	21.097683
814	tired	-20.661144	20.661144

Top 10 F

(g) (GloVe Word Embeddings)Wifi & Connectivity Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
9557	tendency	25.035873	25.035873
9811	tree	24.836399	24.836399
4357	cater	-24.619127	24.619127
5157	martin	23.226828	23.226828
951	attention	-23.201000	23.201000
9954	confuse	22.625372	22.625372
2714	groups	-21.603767	21.603767
7989	angles	-21.533394	21.533394
9765	struggles	21.165909	21.165909
7274	judgement	-21.097683	21.097683

(h) (GloVe Word Embeddings)User Rating Coefficient Analysis.

7.12 All Deep Learning Models FastText pre-trained Word Embeddings Coefficient Analysis

	Feature	Coefficient	AbsoluteCoefficient
7667	niugini	10.781199	10.781199
539	expected	10.712000	10.712000
6844	semi	10.285300	10.285300
8243	skywards	9.885700	9.885700
3959	observed	9.826500	9.826500
934	stand	9.523700	9.523700
6303	promote	9.214200	9.214200
4100	livery	9.120199	9.120199
3844	immaculate	9.018600	9.018600
6969	velocity	8.875800	8.875800

(a) (FastText Word Embeddings)Seat Comfort Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
7665	niugini	10.781199	10.781199
540	flat	10.712000	10.712000
6840	hostage	10.285300	10.285300
8233	skywards	9.885700	9.885700
3956	observed	9.826500	9.826500
933	pacific	9.523700	9.523700
6299	promote	9.214200	9.214200
4098	livery	9.120199	9.120199
3845	immaculate	9.018600	9.018600
6958	velocity	8.875800	8.875800

(b) (FastText Word Embeddings)Cabin Staff Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
7181	balm	10.781199	10.781199
647	children	10.712000	10.712000
7004	awfully	10.285300	10.285300
7835	tullamarine	9.885700	9.885700
3744	bell	9.826500	9.826500
949	swiss	9.523700	9.523700
6584	sk	9.214200	9.214200
4274	ship	9.120199	9.120199
3567	altitude	9.018600	9.018600
9695	propose	8.930400	8.930400

(c) (FastText Word Embeddings)Food & Beverages Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
7486	backward	10.781199	10.781199
372	feel	10.712000	10.712000
9553	registering	10.285300	10.285300
4543	refunding	9.826500	9.826500
782	add	9.523700	9.523700
5423	dramatically	9.214200	9.214200
3987	eligible	9.120199	9.120199
4364	rerouting	9.018600	9.018600
8953	dozed	8.930400	8.930400
6911	carpets	8.546600	8.546600

(d) (FastText Word Embeddings)Ground Service Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
9610	mabuhay	12.102999	12.102999
7774	toasted	10.781199	10.781199
587	previous	10.712000	10.712000
6587	server	10.285300	10.285300
7073	khartoum	9.885700	9.885700
3756	observed	9.826500	9.826500
804	otherwise	9.523700	9.523700
9893	temporarily	9.337299	9.337299
6372	engineering	9.214200	9.214200
4235	nbo	9.120199	9.120199

(e) (FastText Word Embeddings)Inflight Entertainment Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
7692	tasman	10.781199	10.781199
449	upon	10.712000	10.712000
6760	earbuds	10.285300	10.285300
8354	tcx	9.885700	9.885700
3799	dragonair	9.826500	9.826500
830	considering	9.523700	9.523700
6244	pty	9.214200	9.214200
4013	txl	9.120199	9.120199
3875	dimensions	9.018600	9.018600
9716	sights	8.930400	8.930400

(f) (FastText Word Embeddings)Value For Money Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
346	boeing	10.712000	10.712000
4180	abreast	9.826500	9.826500
615	serve	9.523700	9.523700
3989	passangers	9.214200	9.214200
3517	luxurious	9.120199	9.120199
6590	aspire	9.018600	9.018600
5284	pensacola	8.546600	8.546600
2927	global	8.522599	8.522599
8730	airbnb	8.121300	8.121300
4213	registration	-7.836801	7.836801

(g) (FastText Word Embeddings)Wifi & Connectivity Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
7727	tasman	10.781199	10.781199
450	sent	10.712000	10.712000
6801	earbuds	10.285300	10.285300
8384	euroatlantic	9.885700	9.885700
3802	pune	9.826500	9.826500
832	westjet	9.523700	9.523700
6270	pty	9.214200	9.214200
4013	backwards	9.120199	9.120199
3886	dimensions	9.018600	9.018600
9765	struggles	8.930400	8.930400

(h) (FastText Word Embeddings)User Rating Coefficient Analysis.

7.13 Multi Ouput Regression CNN & GRU & LSTM

	Feature	Coefficient	AbsoluteCoefficient
713	mumbai	-12.817374	12.817374
1499	figure	11.659348	11.659348
1616	pushed	-11.497008	11.497008
885	mind	-11.041563	11.041563
934	status	-10.649260	10.649260
1294	august	10.318373	10.318373
3897	ton	10.301453	10.301453
1406	tomorrow	-9.917017	9.917017
6298	disrespectfully	-9.895596	9.895596
4634	peak	9.808111	9.808111

(a) MORS CUstom Word2Vec embeddings Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
8163	astronomical	24.836399	24.836399
3784	goodbye	-24.619127	24.619127
8020	mercedes	-23.308704	23.308704
8407	twisted	23.226828	23.226828
1413	realized	-23.201000	23.201000
2440	assume	-21.603767	21.603767
7241	generation	-21.533394	21.533394
8271	gel	-21.097683	21.097683
909	philadelphia	-20.661144	20.661144
4298	tape	20.640966	20.640966

(b) MORS Glove pre-trained word embeddings Coefficient Analysis.

	Feature	Coefficient	AbsoluteCoefficient
467	compensation	10.712000	10.712000
4081	abreast	9.826500	9.826500
705	compared	9.523700	9.523700
3921	lgw	9.214200	9.214200
3667	cusco	9.120199	9.120199
6411	addons	9.018600	9.018600
38	airline	8.546600	8.546600
2904	segments	8.522599	8.522599
9183	strapped	8.121300	8.121300
4108	registration	-7.836801	7.836801

(c) MORS FastText pre-trained word embeddings Coefficient Analysis.

Figure 7.13: All MORS obtained results for Deep Learning model architectures.

7.14 Multi Output Regression DistilBERT

	Feature	Importance
28176	enoch	62.364281
24991	seamus	61.191628
27869	##may	61.140289
28911	dismissing	60.749485
28989	nano	59.609764
29481	effortlessly	59.218517
29111	##ы	59.132751
28066	creaked	59.009995
25535	ravaged	58.628357
25833	hierarchical	58.494896

(a) MORS pre-trained DistilBERT Coefficient Analysis.

7.15 All Regression Metrics Tables

SVR Obtained Results(TF-IDF)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.98	0.77	0.55
Cabin Staff SRVC	1.19	0.80	0.53
Food & Beverages	1.00	0.77	0.56
Ground Service	0.96	0.73	0.60
Inflight Entertainm	1.14	0.81	0.52
Value For Money	0.87	0.70	0.63
Wifi & Conn	1.18	0.79	0.44
User_Rating	2.65	1.40	0.77

SVR Obtained Results(TF-IDF Multi-Output)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.98	0.78	0.55
Cabin Staff SRVC	1.20	0.86	0.53
Food & Beverages	1.00	0.77	0.56
Ground Service	0.97	0.73	0.60
Inflight Entertainm	1.14	0.81	0.52
Value For Money	0.87	0.70	0.64
Wifi & Conn	1.18	0.79	0.45
User_Rating	2.65	1.22	0.76

Ridge Obtained Results(TF-IDF)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.95	0.78	0.55
Cabin Staff SRVC	1.14	0.85	0.55
Food & Beverages	0.95	0.85	0.57
Ground Service	0.95	0.77	0.60
Inflight Entrt	1.12	0.95	0.53
Value For Money	0.84	0.75	0.65
Wifi & Conn	1.18	0.84	0.44
User_Rating	2.60	1.42	0.76

Ridge Obtained Results(Mulit-Output TF-IDF)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.95	0.79	0.56
Cabin Staff SRVC	1.15	0.86	0.55
Food & Beverages	0.96	0.78	0.57
Ground Service	0.95	0.74	0.61
Inflight Entrt	1.12	0.83	0.53
Value For Money	0.84	0.71	0.65
Wifi & Conn	1.17	0.84	0.45
User_Rating	2.60	1.25	0.77

CNN Obtained Results (GloVe)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.95	0.77	0.52
Cabin Staff SRVC	1.22	0.86	0.50
Food & Beverages	0.95	0.72	0.56
Ground Service	0.99	0.72	0.57
Inflight Entrt	1.48	0.92	0.40
Value For Money	0.86	0.	0.64
Wifi & Conn	1.29	0.77	0.36
User_Rating	3.65	1.37	0.70

CNN Obtained Results (Multi-Output GloVe)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.93	0.76	0.52
Cabin Staff SRVC	1.11	0.54	0.63
Food & Beverages	0.89	0.80	0.67
Ground Service	0.92	0.54	0.74
Inflight Entrt	1.11	0.70	0.45
Value For Money	0.76	0.59	1.01
Wifi & Conn	1.09	0.66	0.81
User_Rating	2.09	0.57	0.80

CNN Obtained Results (FastText)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.94	0.76	0.53
Cabin Staff SRVC	1.04	0.80	0.56
Food & Beverages	1.14	0.83	0.49
Ground Service	1.04	0.71	0.60
Inflight Entrt	1.66	0.99	0.33
Value For Money	0.90	0.72	0.63
Wifi & Conn	1.49	0.81	0.28
User_Rating	3.40	1.32	0.70

CNN Obtained Results (Multi-Output FastText)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.95	0.76	0.52
Cabin Staff SRVC	1.14	0.54	0.64
Food & Beverages	0.91	0.84	0.66
Ground Service	0.91	0.53	0.74
Inflight Entrt	1.12	0.71	0.44
Value For Money	0.79	0.58	1.03
Wifi & Conn	1.11	0.67	0.80
User_Rating	2.20	0.61	0.70

CNN Obtained Results (Custom)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.96	0.76	0.52
Cabin Staff SRVC	1.15	0.80	0.51
Food & Beverages	1.19	0.82	0.47
Ground Service	0.92	0.65	0.63
Inflight Entrt	1.5	0.92	0.39
Value For Money	0.93	0.71	0.62
Wifi & Conn	1.28	0.76	0.38
User_Rating	3.42	1.28	0.72

CNN Obtained Results (Multi-Output Custom)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.89	0.73	0.52
Cabin Staff SRVC	1.05	0.56	0.62
Food & Beverages	0.88	0.79	0.69
Ground Service	0.86	0.57	0.73
Inflight Entrt	1.11	0.68	0.44
Value For Money	0.72	0.59	0.98
Wifi & Conn	1.12	0.65	0.82
User_Rating	1.96	0.63	0.11

Gated Reccurent Unit (GloVe)

Relevant Score	MSE	MAE	R^2
Seat Comfort	1.00	0.75	0.49
Cabin Staff SRVC	1.06	0.73	0.55
Food & Beverages	1.12	0.77	0.50
Ground Service	0.91	0.66	0.63
Inflight Entrt	1.49	0.88	0.39
Value For Money	0.84	0.65	0.66
Wifi & Conn	1.51	0.76	0.31
User_Rating	3.09	1.15	0.73

Gated Reccurent Unit (Multi-Output-GloVe)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.86	0.73	0.54
Cabin Staff SRVC	1.07	0.58	0.63
Food & Beverages	0.87	0.78	0.68
Ground Service	0.87	0.56	0.73
Inflight Entrt	1.06	0.70	0.45
Value For Money	0.74	0.59	0.98
Wifi & Conn	1.08	0.63	0.82
User_Rating	1.98	0.63	0.09

Gated Reccurent Unit (Custom)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.96	0.74	0.52
Cabin Staff SRVC	1.07	0.74	0.55
Food & Beverages	1.16	0.78	0.48
Ground Service	0.88	0.61	0.64
Inflight Entrt	1.48	0.87	0.40
Value For Money	0.83	0.65	0.66
Wifi & Conn	1.31	0.79	0.34
User_Rating	3.09	1.15	0.73

Gated Reccurent Unit (Multi-Output Custom)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.83	0.70	0.56
Cabin Staff SRVC	1.02	0.59	0.59
Food & Beverages	0.80	0.75	0.71
Ground Service	0.83	0.58	0.72
Inflight Entrt	1.01	0.68	0.47
Value For Money	0.68	0.63	0.91
Wifi & Conn	1.03	0.65	0.83
User_Rating	1.83	0.64	0.11

Gated Reccurent Unit (FastText)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.89	0.71	0.55
Cabin Staff SRVC	0.96	0.72	0.61
Food & Beverages	1.01	0.74	0.55
Ground Service	0.82	0.60	0.68
Inflight Entrt	1.46	0.86	0.44
Value For Money	0.79	0.64	0.68
Wifi & Conn	1.19	0.76	0.40
User_Rating	3.09	1.15	0.73

Gated Reccurent Unit (Multi-Output-FastText)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.92	0.80	0.53
Cabin Staff SRVC	1.10	0.55	0.66
Food & Beverages	0.86	0.85	0.69
Ground Service	0.87	0.54	0.79
Inflight Entrt	1.08	0.75	0.45
Value For Money	0.71	0.60	0.96
Wifi & Conn	1.08	0.73	0.84
User_Rating	1.74	0.62	0.10

LSTM Obtained Results(GloVe)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.98	0.76	0.52
Cabin Staff SRVC	0.74	0.72	0.63
Food & Beverages	0.94	0.73	0.57
Ground Service	0.79	0.63	0.68
Inflight Entrt	1.2	0.84	0.49
Value For Money	0.77	0.65	0.69
Wifi & Conn	1.21	0.75	0.40
User_Rating	1.93	0.96	0.82

LSTM Obtained Results(Multi-Output GloVe)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.91	0.76	0.51
Cabin Staff SRVC	1.11	0.55	0.60
Food & Beverages	0.89	0.83	0.70
Ground Service	0.90	0.54	0.73
Inflight Entrt	1.14	0.72	0.45
Value For Money	0.70	0.59	0.89
Wifi & Conn	1.10	0.66	0.83
User_Rating	1.85	0.61	0.11

LSTM Obtained Results(FastText)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.81	0.71	0.60
Cabin Staff SRVC	0.87	0.72	0.64
Food & Beverages	0.95	0.75	0.58
Ground Service	0.78	0.67	0.69
Inflight Entrt	1.32	0.90	0.47
Value For Money	0.73	0.65	0.70
Wifi & Conn	1.21	0.79	0.48
User_Rating	2.40	1.10	0.78

LSTM Obtained Results(Multi-Output FastText)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.97	0.81	0.49
Cabin Staff SRVC	1.13	0.52	0.65
Food & Beverages	0.94	0.86	0.68
Ground Service	0.93	0.54	0.77
Inflight Entrt	1.20	0.76	0.44
Value For Money	0.75	0.57	0.99
Wifi & Conn	1.12	0.71	0.81
User_Rating	2.03	0.60	0.12

LSTM Obtained Results(Custom)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.79	0.72	0.59
Cabin Staff SRVC	0.80	0.73	0.62
Food & Beverages	0.95	0.74	0.59
Ground Service	0.77	0.65	0.68
Inflight Entrt	1.26	0.86	0.5
Value For Money	0.75	0.65	0.68
Wifi & Conn	1.18	0.77	0.45
User_Rating			

LSTM Obtained Results(Multi-Output Custom)

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.90	0.75	0.51
Cabin Staff SRVC	1.09	0.56	0.59
Food & Beverages	0.88	0.83	0.71
Ground Service	0.89	0.55	0.74
Inflight Entrt	1.13	0.71	0.45
Value For Money	0.67	0.59	0.90
Wifi & Conn	1.09	0.67	0.83
User_Rating	1.80	0.62	0.13

DistilBERT Obtained Results

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.89	0.72	0.58
Cabin Staff SRVC	0.96	0.67	0.62
Food & Beverages	1.05	0.75	0.56
Ground Service	0.82	0.60	0.69
Inflight Entrt	1.30	0.81	0.50
Value For Money	0.74	0.62	0.71
Wifi & Conn	1.14	0.71	0.48

DistilBERT Multi Output Obtained Results

Relevant Score	MSE	MAE	R^2
Seat Comfort	0.83	0.70	0.51
Cabin Staff SRVC	1.01	0.59	0.58
Food & Beverages	0.82	0.73	0.72
Ground Service	0.79	0.57	0.77
Inflight Entrt	1.12	0.67	0.42
Value For Money	0.64	0.60	0.83
Wifi & Conn	1.14	0.61	0.86
User_Rating	1.40	0.62	0.13