UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Analyzing Customer Reviews on Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence (XAI)

by Anirban Adak

Thesis submitted in fulfilment of the requirements for the degree of

Master in Science (Research) in Computing Sciences

under the supervision of Distinguished Professor Biswajeet Pradhan and Dr Nagesh Shukla

University of Technology Sydney
Faculty of Engineering and Information Technology

November 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, **Anirban Adak** declare that this thesis, is submitted in fulfilment of the requirements for the award of **Master of Science (Research) in Computing Sciences**, in the **Faculty of Engineering & Information Technology** at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature:

Production Note:
Signature removed prior to publication.

Anirban Adak

Date: 21 November 2022

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and respect to my esteemed principal supervisor, Distinguished Professor Biswajeet Pradhan and my co-supervisor Dr Nagesh Shukla. Their achievements in scientific world has been a continuous inspiration for me. Their continuous guidance throughout the research duration, all the way from assessing the research questions, defining research methodology, validating the experiments, critical reviews, and suggestions helped me immensely. Each review comment backed up by so much experience was always fascinating to see how it was changing the direction or flow of the research, and scientific papers that we published during my research period. Without their constant encouragement and supervision, this research would not have been possible.

I would like to extend my sincere gratitude to the candidature assessment panellists who patiently listened to my research and provided valuable feedback to shape up the research.

I am very grateful to the Centre for Advanced Modelling and Geospatial Information Systems, Faculty of Engineering and Information Technology, the University of Technology Sydney for offering research scholarships that have enabled me to accomplish this study.

I express my deep gratitude to my parents for their continuous encouragement and my wife for letting me spend as much time as possible on my study.

I also thank everyone whose name is not included here but have helped me directly or indirectly.

LIST OF PAPERS/PUBLICATIONS

Following papers are produced as a part of the research:

- Adak, Anirban, Biswajeet Pradhan, and Nagesh Shukla. "Sentiment Analysis of Customer Reviews of Food Delivery Services Using Deep Learning and Explainable Artificial Intelligence: Systematic Review." *Foods 11*, no. 10 (**2022**): 1500 – **Published**

- Adak, Anirban, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri. 2022. "Unboxing Deep Learning Model of Food Delivery Service Reviews Using Explainable Artificial Intelligence (XAI) Technique", *Foods 11*, no. 14 (**2022**): 2019 – **Published**

Note: Thesis includes the contents from the papers published above.

TABLE OF CONTENTS

CERTIFICATE OF ORIGINAL AUTHORSHIP ..... **i**

ACKNOWLEDGEMENT ..... **ii**

LIST OF PAPERS/PUBLICATIONS ..... **iii**

LIST OF TABLES ..... **vii**

LIST OF FIGURES ..... **viii**

LIST OF ABBREVIATIONS ..... **ix**

ABSTRACT ..... **x**

1 INTRODUCTION **1**

1.1 General Introduction 1
1.2 Research Background 4
1.3 Research Gaps 5
1.4 Scope of Thesis 6
1.5 Motivation behind Research 7
1.6 Research Aim and Objectives 8
1.6.1 Objective 1 9
1.6.2 Objective 2 9
1.6.3 Objective 3 9
1.7 Research Questions 10
1.7.1 Questions pertaining to objective 1 10
1.7.2 Questions pertaining to objective 2 10
1.7.3 Questions pertaining to objective 3 10
1.8 Research Hypothesis 10
1.8.1 Hypothesis 1 10
1.8.2 Hypothesis 2 11
1.8.3 Hypothesis 3 11
1.8.4 Hypothesis 4 11
1.9 Novelty and Main Contribution 12
1.10 Thesis Organisation 14

2 LITERATURE REVIEW **15**

2.1 Introduction 15

2.2 Literature Review Methodology 16
2.2.1 Aim and Research Questions 18
2.2.2 Search and Selection Process 18
2.3 Previous work on FDS using sentiment analysis 20
2.3.1 Traditional approaches on FDS using sentiment analysis 20
2.3.2 Machine learning approaches on FDS using sentiment analysis 21
2.3.3 Explainable AI techniques 26
2.3.4 Topic Categorization 33
2.4 Strength and limitations of models 35
2.5 Current research issues in food delivery services 37
2.6 Summary 37

3 MATERIALS AND RESEARCH METHODOLOGY 39

3.1 Introduction 39
3.2 Data Acquisition 39
3.2.1 Data scraping using ParseHub 39
3.2.2 Identify Data Attributes 41
3.2.3 Data Splitting 43
3.2.4 Data cleansing 43
3.3 RNN Architecture 43
3.4 Deep learning techniques 45
3.4.1 LSTM and Bi-LSTM 45
3.4.2 Bidirectional GRU 46
3.5 XAI Techniques 46
3.5.1 SHAP 46
3.5.2 LIME 47
3.6 LDA 48
3.6.1 Methods for finding the optimal number of topics in LDA 48
3.7 Overall Methodology 49
3.8 Implementation of the methodology 52
3.8.1 Objective 1 52
3.8.2 Objective 2 54
3.8.3 Objective 3 55
3.8.4 Evaluation and performance metrics 56
3.9 Summary 58

4 RESULTS AND DISCUSSION 60

4.1 Introduction 60
4.2 Results of Objective 1 60
4.2.1 Sentiment Analysis using simple and Hybrid DL models 60
4.2.2 Discussion 60
4.2.3 Validation 65
4.3 Results of Objective 2 66
4.3.1 XAI explanation on LSTM model using SHAP and LIME 66
4.3.2 Discussion 66
4.3.3 Validation 69
4.4 Results of Objective 3 69
4.4.1 Topic Categorization of negative and positive sentiments using LDA 69
4.4.2 Discussion 70
4.4.3 Validation 81
4.5 Summary 81

5 CONCLUSIONS AND FUTURE WORK RECOMMENDATIONS 83

5.1 General Conclusion 83
5.2 Conclusion of Objective 1 83
5.3 Conclusion of Objective 2 84
5.4 Conclusion of Objective 3 85
5.5 Research Drawbacks and Limitations 86
5.6 Recommendations for Future Work 86

REFERENCES 87

LIST OF TABLES

Table 2.1. Search queries and results showing the number of papers. 19
Table 2.2. Literature classification. 19
Table 2.3. Interpretability of methods used for sentiment analysis in FDS. 27
Table 3.1. Different attributes of the dataset from ProductReview. 41
Table 3.2. Confusion Matrix 56
Table 4.1. Performance metrics - (a) LSTM; (b) Bi-LSTM; and (c) Bi-GRU-LSTM-CNN model. 64
Table 4.2. Accuracy scores achieved in ML/DL models from recent papers. 65
Table 4.3. Word contribution for topic on (a) negative and (b) positive reviews. 70
Table 4.4. Category Names derived from Keywords with weights 74
Table 4.5. Coherence score and perplexity for no. of topics. 77
Table 4.6. Positive and Negative Categories extracted from customer reviews 81

LIST OF FIGURES

Figure 1.1. High-level AI diagram. 3
Figure 1.2. Solution framework for sentiment analysis in FDS. 13
Figure 2.1. Literature review methodology 17
Figure 2.2. Classifications of techniques for Sentiment Analysis. 20
Figure 3.1. ProductReview website for Menulog (www.productreview.com.au) 40
Figure 3.2. Wordcloud of customer reviews from productreview site. 41
Figure 3.3. Negative and positive sentiment count. 43
Figure 3.4. Showing RNN architecture. 44
Figure 3.5. LSTM architecture 46
Figure 3.6. Dependencies in LDA. 48
Figure 3.7. Overall Methodology flow chart with DL model, XAI technique and LDA model adopted in this work. 51
Figure 3.8. Methodology flow chart with DL technique adopted in this work. 52
Figure 3.9. Methodology flow chart with XAI technique adopted in this work. 54
Figure 3.10. Methodology flow chart with LDA adopted in this work. 55
Figure 4.1. SHAP explanation on the positive customer review. 67
Figure 4.2. SHAP explanation on the negative customer review. 67
Figure 4.3. LIME explanation on the positive customer review detected by the LSTM model. 68
Figure 4.4. LIME explanation on the negative customer review detected by the LSTM model. 68
Figure 4.5. Coherence score vs no. of topics on (a) negative (b) positive reviews. 76
Figure 4.6. (a) (b) (c) Topics with keywords for negative sentiments. 79
Figure 4.7. (a) (b) Topics with keywords for positive sentiments. 80

LIST OF ABBREVIATIONS

BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU-LSTM-CNN	Embedded Bidirectional GRU LSTM CNN
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
DL	Deep Learning
DT	Decision Trees
FDS	Food Delivery Services
LIME	Local Interpretable Model Agnostic Explanation
LDA	Latent Dirichlet Allocation
LSTM	Long short-term memory
ML	Machine Learning
NB	Nave Bayes
NN	Neural Networks
OA	Overall Accuracy
SHAP	Shapley Additive Explanations
SVM	Support Vector Machines
XAI	Explainable Artificial Intelligence

ABSTRACT

Social media reviews and feedback are getting increasingly important for customers ordering food from a food delivery services in the last few years. This trend has become even more prominent since COVID-19 pandemic and government enforced lockdowns. During the Covid-19 crisis, customer's preferences in having food delivered to their doorstep instead of waiting in a restaurant has propelled the growth of food delivery services (FDS). As all restaurants go online and get onboarded to FDS, such as UberEATS, Menulog or Deliveroo, customer review on online platforms has become an important source of information about the company's performance. The FDS organisations would like to find complaints from customer feedback and use the data effectively to understand the areas for improvement to enhance customer satisfaction. The study aims to review the Machine Learning (ML) and Deep Learning (DL) models along with explainable artificial intelligence (XAI) method to predict customer sentiment in the FDS domain. This research aims to develop a robust end-to-end framework using AI/ML which can help to accurately predict customer sentiment in the first objective. The second objective presents the XAI technique implementation on the black box DL models. The explanations of the black box models as how they build the outcome will help build the trust in the system. The third objective groups the positive and negative sentiments in groups using topic categorization technique. The groups can be used for sending the customer complaints for process improvement and positive reviews for rewarding staff. Firstly, in the objective 1, customer review data was collected from Productreview website and was used for building simple Long short-term memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM) and hybrid Embedded Bidirectional GRU LSTM CNN (Bi-GRU-LSTM-CNN) DL models for performing sentiment analysis. The DL models were compared to pick the best classifier for FDS domain. The results showed LSTM model, Bi-LSTM model and Bi-GRU-LSTM-CNN model achieved accuracy of 96.07%, 95.85% and 96.33% respectively. Secondly, in the objective 2, XAI techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model Agnostic (LIME) were used on the best DL model to provide explanation on the sentiment prediction. Both the techniques SHAP and LIME proved useful in explaining the model with features (words in case of sentences) which are contributing the prediction outcome.

Thirdly, in the objective 3, this study implemented topic categorization technique LDA on the positive and negative comments.

Keywords: sentiment analysis, food delivery services, deep learning, explainable artificial intelligence, lime, shapley

1 INTRODUCTION

This chapter reflects a general introduction, research background along with customer sentiment analysis in Food Delivery Services (FDS) and unboxing deep learning models using Explainable Artificial Intelligence (XAI). This chapter also reveals the main context of the study, structure, problem statement, specified objectives, research goal, research plan, specific research questions, motivation, research limitation, and thesis organization. It highlights the benefits of sentiment analysis which will enable FDS organisation to identify and resolve customer negative reviews, which will in turn increase customer satisfaction.

1.1 General Introduction

Customer satisfaction is the key to assess how a product or service of a company meets customer expectations (Kefa and Kendi 2019). It is an important tool that can give organisations major insights into every part of their business, helping them to earn more money or minimise marketing expenses (Barsky and Labagh 1992). Customer feedback might help in reviewing the factors that were not previously considered, such as shipping, safe packing, politeness and available customer service consultants, a user-friendly website and others. Nothing can make customers feel that they are more important than asking for their views and taking their comments seriously. When a customer is asked for any opinion on a product or experience, they feel valued and connected to the organisation (Suhartanto et al. 2019). In the food industry, customers often look into restaurant reviews before placing their orders. These days, restaurants or food delivery services (FDS) have a review or feedback system that is integrated with their portal or social media platforms, but very few act on customer opinions. This situation can be due to the presence of a large amount of review data across various platforms and lack of customer service consultants to go through each of them to act on it (Ara et al. 2020). At present, organisations need not depend any-more on customer service consultants to read all the reviews. Instead, organisations can rely on artificial intelligence (AI) to solve their problems and save costs.

Post Covid 19 pandemic, with the rise of online food delivery marketplaces, FDS have brought versatility and a variety of restaurants to the comfort and convenience of homes and offices (Parliament of Australia 2018). In addition, an increase in immigration from different countries has given rise to new cuisines being introduced into the country. Customers are provided a wide range of meal options and the ability to order from the best eateries or restaurants in town, all done from their home or office. With applications becoming a standard utility on mobile devices and global positioning systems (GPS) made available to all, the delivery of food to a customer's exact location is no longer an issue. Customers can track the progress of their order from the time of order until it arrives at their door. With the rising demand for food takeaway services, more digital market-place platforms are jumping on the bandwagon.

Globally, ordering and delivery marketplace platforms such as UberEATS, Deliveroo and Menulog (Sue 2018) operate in a more cost-intensive business model but take responsibility for the entire delivery logistics. These companies offer a complete sales solution to the restaurants and food business owners at no extra cost and work on a commission-based model. With a few taps on the phone by the customer, FDS applications receive orders, pick up the food from restaurants and deliver it to the customer. Customers have various food options from a chain of restaurants. Online food companies are delighted to find out that customers are eager for such services. Amidst projections that Australia's food delivery industry would grow (Statista 2021), Covid-19 lockdowns and quarantines have led to an increase in FDS (Reiley 2020) including the use of third-party apps such as UberEATS, Deliveroo and Menulog, as more people are forced to order online while restaurants are closed. With more orders and feedbacks, most of the companies want to use the data effectively to understand the areas for improvement to enhance the customer satisfaction.

The use of AI in natural language processing (NLP) has immense potential to determine the positive, negative and neutral reviews (Geler et al. 2021). Machine Learning (ML) and Deep Learning (DL) are often used interchangeably in AI but have different meanings. At a high level, ML is the method of data analysis that automates analytical model building, whereas DL is the subset of ML (see Figure 1.1) concerned with

algorithms inspired by the structure and function of the brain called artificial neural networks (LeCun, Bengio, and Hinton 2015).

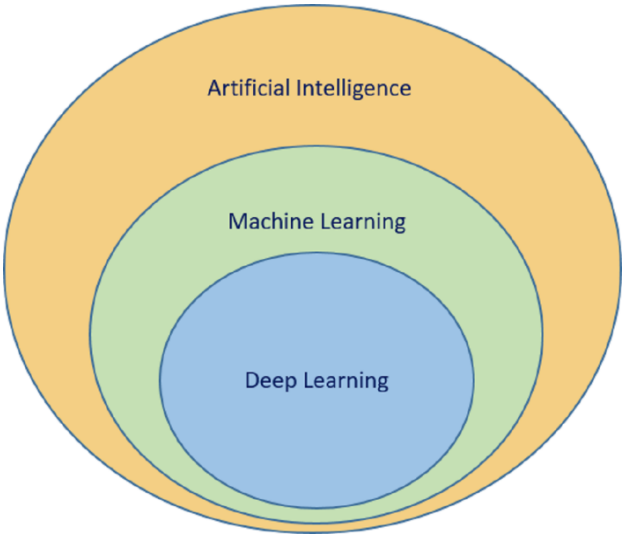


Figure 1.1. High-level AI diagram.

Customer sentiment can be found in blog posts, comments, reviews or tweets that mention the quality of food, service, delivery time and other details (Lokeshkumar et al. 2020). The FDS organisations can understand what customers are saying and perceive positive comments as compliment and negative comments as complaints (Singh and Verma 2020). The negative sentiments can be classified into various complaint categories using topic modelling. Customer experience with food can vary with different seasons as positive feed-back increase during the peak season (Yu and Zhang 2020). Despite huge revenues and investments, FDS organisations still struggle with profitability due to high expenses. Predatory pricing is a commonly used strategy to beat the competitive market where businesses swallow a sales loss by massively subsidising meal costs. Furthermore, online FDS have minimal control over food quality as it is more dependent on the restaurants. If a customer is dissatisfied with the quality, the food delivery company needs to bear the loss of revenue. As a result, businesses such as Sprig (Failory.com 2017) and Munchery (Techcrunch.com 2019) are unable to endure the loss of revenue and have exited the business (Jiang 2020). The only way for food delivery companies to ensure that the customer experience of the delivery operation is good and does not damage the dine-in experience is by tracking customer reviews and feedbacks.

By realizing the importance of customer feedback, complexity with large volume of customer review data and success of artificial intelligence in other fields to improve prediction accuracy, FDS organisations can automate the process of predicting customer sentiment and work towards improving the issues. There is always a trade-off between accuracy and interpretability of the while selecting machine learning models. The black-box deep learning models produce high accuracy but often lags on interpretability due to which it is difficult to explain the rationale of behind the decisions made. Explainable artificial intelligence (XAI) promises to resolve the issue of explainability and interpretability of DL black boxes (Lorente et al. 2021).

1.2 Research Background

Customer management, an important factor in the FDS business, is measured with customer engagement. Retaining customers becomes extremely crucial when the market is competitive and one desires to improve on the FDS (Upadhyay et al. 2022). The first step in customer engagement is to receive feedback and reviews. Feedback acts as a learning tool that makes customers feel important and valued. One needs to rectify their limitations for an enhanced takeaway home delivery system, and that is only possible by analysing genuine feedback from customers. Sentiment analysis is a form of information that comes directly from the customers about their overall experience and opinion about a business, product or service (Akila et al. 2020). The experience can be in the form of satisfaction or dissatisfaction, and may be positive, negative or neutral (Akila, et al. 2020). Sentiment analysis, also known as opinion mining, has gained significant importance over time due to the steep increase in the amount of customer feedback available online in the form of tweets or reviews (Nagpal et al. 2020). People share their opinions on restaurants and food on social media and make their comments visible to any person on the internet. Feedback helps customers to decide on product purchases. More positive feedbacks from customers increases the chances of selling the product and attract more attention in the market. Sentiment analysis is important for businesses and decision-makers (Akila et al. 2020) as it helps in getting market insights that help companies to identify the key areas in improving customer experience and enhancing their brands. Today, when a customer orders food online using websites or mobile apps, a pop-up

window appears asking for feedback, which greatly increases customer engagement. When customers plan to order food online, they would prefer to look for accurate reports (Singh and Verma 2020). If the FDS app or website does not have any online reviews, customers may change their decision to order. Having 'no reviews' can be just as detrimental as having negative reviews. Having genuine and positive reviews helps to increase the credibility factor. Negative reviews are difficult to handle for any business. They can take away the potential number of customers away from the FDS and cause the existing customers to question whether they want to re-order. Thus, FDS operators have to remember that they cannot control every customer's experience, mistake or circumstance. On the bright side, a negative review can provide insights into weaknesses and provide opportunities to improve the customer service (Lan et al. 2016).

The key benefits of sentiment analysis (Nagpal et al. 2020) on business are the following:

- keeps businesses connected round the clock with the customers;
- provides business insights to help in decision-making;
- indicates real-time trends with emotion data;
- helps improve the business plan of action to gain an advantage over competitors;
- can be conducted on services or products to understand which item is eliciting negative sentiments;
- provides a great tool for businesses to improve customer service in any domain.

1.3 Research Gaps

Several papers (Rai and Shukla 2022; Akila et al. 2020; Nagpal et al. 2020; Lan et al. 2016) reveals, previous research works were done based on text analysis over supply chain domain using different machine learning algorithms to understand customer emotions, but they fail to classify the negative sentiments of the customers into something useful and solve the supply chain problems. Due to competitive market and rise in sales every year, food delivery service companies would like to get data classified as why customer gave the negative feedbacks and what measures could be taken to improve customer service. Also, it is a gigantic task for the food delivery service companies to keep one person to read, understand and redirect the review/feedback to appropriate

solution areas. There are deep learning techniques which can increase the performance and accuracy for identifying the issues. With the continuously increasing volume of customer review data, a robust end-to-end framework using AI/ML can help accurately predict customer sentiment. Such a framework will be beneficial for FDS organisations, such as Ubereats, Menulog and Deliveroo. However, deep learning models have the drawback of not being human interpretable model raising concerns about model's interpretability. Zucco et al. (2018) revealed very few works have been done to explain its decision-making process and actions. Also, So (2020) expressed the need for uncovering the machine learning models using XAI which is used in sentiment analysis of customer reviews for hotel. Although XAI techniques are recommended for examining deep learning models in other industries. However, to the best of our knowledge, there is no evidence of application of XAI techniques along on DL models in the FDS industry to analyse customer reviews.

Therefore, in this section, we highlight the main research gaps obtained from an extensive literature review are:

- Earlier research work does sentiment analysis of the reviews without looking for the solutions to improve customer satisfaction.
- Lack of explainability of the black box algorithms is an issue for the industry to trust the deep learning techniques.
- Because of rise of volume of feedbacks as review data, it is better to use Artificial Intelligence to classify and categorize the data instead of doing it manually.

1.4 Scope of Thesis

Customer’s preference for having food delivered to their doorstep rather than waiting in a restaurant propelled the growth of FDS during the COVID-19 crisis. Customer reviews on online platforms have become an important source of information about a company’s performance, with all restaurants going online and bringing FDSs onboard, such as UberEATS, Menulog or Deliveroo. FDS organisations strive to collect customer complaints and effectively use the data to identify areas for improvement in order to improve customer satisfaction. However, due to large customer feedback data and lack

of customer service consultants, only a few customer opinions are addressed. Organizations can use AI to solve problems and save money instead of relying on customer service consultants to read all of the reviews. Based on literature, deep learning (DL) methods have shown remarkable results in getting better accuracy working with large datasets in other domains but lacks in explanation of its model. Rapid research on XAI to explain predictions made by opaque models looks promising and it is yet to be researched in FDS domain. Thus sentiment analysis on customer reviews using DL models, implementation of XAI techniques to analyse DL model's logic and topic modelling of the sentiments are the major scopes.

Thus the scope of this research deals with:

- Sentiment analysis of customer reviews using DL models
- Implementation of XAI techniques to analyse the DL model's prediction logic.
- Perform topic modelling on the negative and positive customer reviews.

In this thesis, customer review data was collected from the Product Review website which was later used for training and testing DL models. The scope of this research is more towards developing and comparing DL models with explainability. The DL models were developed based on LSTM, Bi-LSTM and Bi-GRU-LSTM-CNN algorithms. The accuracy of the models was accessed based on the precision, recall and confusion matrix metrics. The DL model with lesser false negatives is picked as the best model for performing sentiment analysis. Different XAI techniques like SHAP and LIME are implemented on the best model to understand the prediction logic. Finally, topic modelling is done on the customer reviews to determine various topics which can be used to categorize and sent to the concerned department which can work on service improvement.

1.5 Motivation behind Research

COVID-19 lockdowns and quarantine have increased the demand for online food delivery service (FDS) organisations such as Ubereats, Deliveroo and Menulog, as restaurants were instructed to stop dining services (Laguna et al. 2020; Poelman et al. 2021). When customers order cuisines from online food delivery services, they primarily look for

reviews and recommendations of others. Positive reviews drive customers to take decision to order food from one restaurant, whereas negative reviews can help to look for other options. Food delivery service companies can look for the negative comments towards common com-plaint types such as customer service, food quality, cost or slow delivery service to understand the improvement areas to enhance the customer satisfaction. Restaurants and FDSs now have a review or feedback system integrated into their portals or social media platforms; however, due to the large amount of review data spread across multiple platforms and the lack of customer service consultants to go through and act on each of these comments, only a few respond to customer feedback (Ara et al. 2020). Organizations no longer need to hire customer service consultants to read all of the reviews since artificial intelligence (AI) can help them solve problems and save money (Mhlanga 2018; Panda et al. 2019).

Thus motivation of this thesis indicates that in FDS domain due to cutthroat competition in the FDS industry, FDS organisations want to improve customer satisfaction by acting on the customer complaints. But due to large amount of customer review data, it is difficult for organisations to go through each one of them to find the complaints. With help of AI, the negative and positive sentiment reviews can be easily classified from large volumes of customer review data. The negative customer review can be used for improving customer satisfaction whereas positive customer review can be used for rewarding staff and restaurants. This research is more focussed on increasing the accuracy of the model along with prediction logic of the outcome which can be verified.

1.6 Research Aim and Objectives

The aim of the study is to develop DL models for performing sentiment analysis of customer reviews along with topic modelling of the sentiments in FDS domain. Also, the DL models are examined using XAI technique for validating the prediction logic.

The current research developed three DL models that fulfil the research gap in the literature. The proposed models are complex and opaque in nature (Luo and Xu 2021). The best fit DL model's prediction was verified using XAI technique for the predicting

outcome. The predicted positive and negative sentiments were grouped using topic modelling. The main objectives of the present research are as follows:

- To develop effective deep learning models to do sentiment analysis of the customer reviews.
- To develop an effective explainable AI technique interoperable to the deep learning model to produce AI model explanation.
- To develop technique to perform topic categorization on the negative sentiments found on the previous step and then propose solution related to it.

1.6.1 Objective 1

The first objective of the designed approach is to develop different DL models using customer review dataset collected from the Product Review website. The raw data from the website would be cleaned to remove noise before using it for training and testing the DL models. The best fit DL model can be used to analyse customer sentiments as positive or negative sentiments.

1.6.2 Objective 2

The second objective of the research is to assess the DL model for its prediction logic using XAI techniques. Two different XAI techniques can be implemented on the DL model to understand the features which are contributing the outcome. The words which are contributing negative or positive sentiment can be verified if they are correct in doing so. This will give trust to the businesses to use the black box DL models for analysing customer reviews.

1.6.3 Objective 3

The third objective is to use topic modelling technique to find various topics from the customer reviews. This can help the business to group the negative and positive sentiments based on the topic and assign it to the concerned department to fix the issue. This can help the FDS organisations to improve the supply chain issues.

The overall designed framework has novelty and prepared based on analysing various older and recent models for analysing customer reviews in FDS domain. The framework

developed in the current research including DL models, XAI techniques along with topic modelling technique makes it comprehensive and accurate.

1.7 Research Questions

The research aims of this paper is to address the gap identified in the literature by answering the following research questions:

1.7.1 Questions pertaining to objective 1

Which deep learning classifier would be best suited to pick FDS customer complaints from feedback and work on its solution?

1.7.2 Questions pertaining to objective 2

Can XAI techniques like LIME or SHAP provide explanation on sentiment prediction and build trust on the deep learning model created from the previous question?

1.7.3 Questions pertaining to objective 3

Can topic modelling technique like LDA find various topics from the customer reviews which can be used by FDS organizations to analyse the real problems and send it to concerned department to resolve customer issues?

1.8 Research Hypothesis

The research hypothesis tries to look at the research questions to find the various scenarios of the research outcomes. The following hypothesis are expected while undergoing the research work.

1.8.1 Hypothesis 1

Expectation - DL models would give higher accuracy with no explainability while performing sentiment analysis on the customer reviews in FDS domain.

Testable - The accuracy of the DL models can be found using different performance metrics.

Falsifiable - If the accuracy of the DL models come below minimum expectation of 90%, then the hypothesis of DL models having high accuracy over other ML models would fail.

1.8.2 Hypothesis 2

Expectation - The interpretability of the black box DL models can be brought using XAI techniques.

Testable – The feature contribution of the words contributing to positive or negative biasness of the sentence can explain the interpretability of the DL models. The prediction logic of the DL models can be tested using XAI.

Falsifiable – If the prediction logic of the highly accurate DL models is not found using XAI methods, then the hypothesis that XAI can bring interpretability of the black box models would fail.

1.8.3 Hypothesis 3

Expectation - The negative customer reviews found from DL models can be categorised using Topic Modelling technique to classify the problems and issues which can be solved by FDS organisations.

Testable – The topics generated using the words used in the sentences can be found using Topic Modelling techniques. The topic modelling method should generate many topics based on the negative review corpus/dataset.

Falsifiable – If the topic modelling method is not able to classify the sentiments into meaningful categories based on the keywords used, then the hypothesis that Topic Modelling technique can categorize review dataset into meaningful categories would fail. FDS organisations then have to manually read the negative comments and solve customer problems.

1.8.4 Hypothesis 4

Expectation - The positive customer reviews found from DL models can be categorised using Topic Modelling technique to reward the staff and restaurants.

Testable – The topics generated using the words used in the sentences can be found using Topic Modelling techniques. The topic modelling method should generate many topics based on the positive review corpus/dataset.

Falsifiable – If the topic modelling method is not able to classify the sentiments into meaningful categories based on the keywords used, then the hypothesis that Topic Modelling technique can categorize review dataset into meaningful categories would fail. FDS organisations then have to manually read the positive comments and find the appreciation made in the review dataset.

1.9 Novelty and Main Contribution

The main novelty and contribution of the research work is towards the building deep learning model and then explaining the model using the XAI technique to validate the model's logic for food industries to use it. Furthermore, topic modelling helps in identifying the areas for improvement for FDS organisations. Based on the recommendations and gaps found in the previous section, we designed a robust end to end framework using AI/ML can help accurately predict customer sentiment. Below is the diagram Figure 1.2, which consist of the solution framework which consists of DL model, XAI methods and topic categorization.

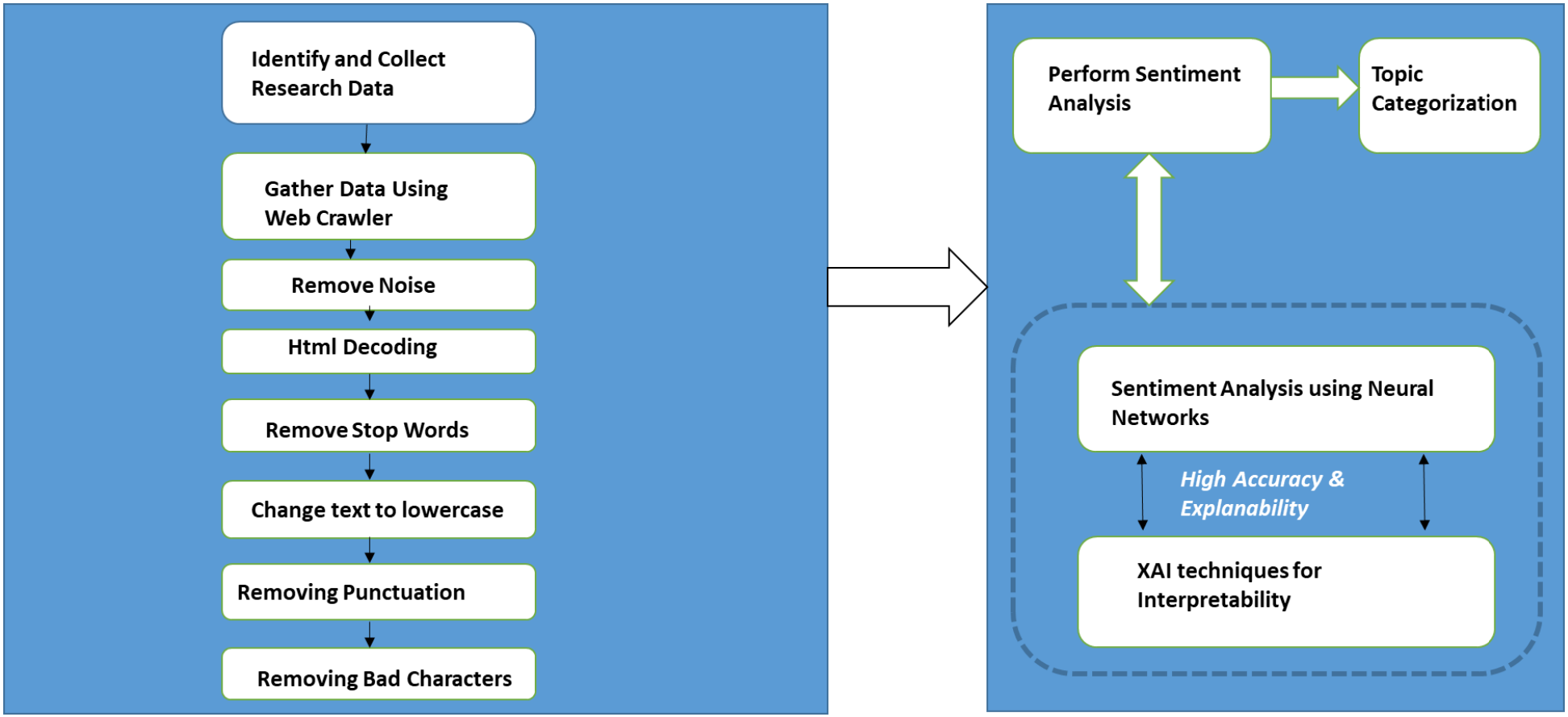


Figure 1.2. Solution framework for sentiment analysis in FDS.

1.10 Thesis Organisation

The thesis consists of five chapters. The detail of contents carried out by the chapters were pointed out below.

Chapter 1 reveals the introduction to the topic and research background, sentiment analysis in FDS, research problem, research gap, aim of the research, objectives and questions, the scope of the study, motivation behind this research, novelty and main contribution of the research and thesis organization in detail.

Chapter 2 demonstrates the literature on sentiment analysis performed in FDS and other domains. The first part of the chapter mainly discusses about the previous work done and the methodology of various models used for performing sentiment analysis on customer reviews. In the second part, comparative analysis in terms of the limitations and strengths of the models are discussed. Furthermore, the paper highlights the research issues in FDS.

Chapter 3 in the thesis discusses the methodology and the proposed models. This chapter demonstrates and discusses the data acquisition, study area, overall methodology, and implementation of the developed models for sentiment analysis.

Chapter 4 describes the accuracy results of DL models, explanation of the best DL model using XAI method and topic categorization using topic modelling.

Chapter 5 concludes the study with detail description of research limitations, main findings, and future directions.

2 LITERATURE REVIEW

This chapter provides an extensive review of several traditional and machine learning techniques for performing sentiment analysis on customer reviews. The traditional models and machine learning are discussed by highlighting the involvement of some supporting models, uncertainties, and accuracy. This chapter also presented a comprehensive review of deep learning and explainable AI techniques used in FDS and other domains. Also, the chapter discusses on the methods used for topic modelling in different domains. Finally, the strength and limitations of the models are discussed. The last section highlights the current research issues in food delivery services faced today. In general, this chapter reflects a general view of the use of several models for sentiment analysis of the customer reviews.

2.1 Introduction

In literature, many papers (Upadhyay et al 2022; Akila et al 2020; Nagpal et al. 2020; Lan et al 2016) have presented various models for performing the sentiment analysis of customer review in FDS domain. The solution to predict the sentiment of customer reviews in FDS domain has evolved from lexicon methods to ML and DL. Lexicon-based systems are quick to train, but ML-based systems achieve state-of-the-art sentiment analysis performance. Due to the high complexity of hybrid-based proposals, they are not yet widely used. As a result, ML-based approaches are without a doubt the most popular for sentiment analysis, with models such as Nave Bayes (NB), Maximum Entropy (ME), Decision Trees (DT), Support Vector Machines (SVM) and Neural Networks (NN) being frequently used in the literature (Mabrouk et al 2020). In fact, the last one, NN, is widely used (Moraes et al 2013; Zhang et al 2015; Tang, et al 2015; Zhang et al 2018) due to its superior efficiency (high performance and quick execution) over the other options.

2.2 Literature Review Methodology

A standard review process can be described in three steps: plan, conduct and report (Kitchenham and Charters 2007).

Step 1: Review planning, which is crucial due to the following reasons:

- COVID-19 has increased the demand for online FDSs;
- Improving customer satisfaction and meeting customer expectations;

Challenges in the adaptation of DL methods for sentiment analysis due to the reduced explainability of models.

The first step was divided into various sections such as 'Aim and research question', 'Search and selection process', 'Inclusion and exclusion criteria', 'Quality assessment' and 'Data extraction and synthesis'.

Step 2: A review phase was conducted by searching and identifying relevant journals and articles with the following keywords: 'sentiment analysis of customer reviews', 'food', 'deep learning', 'machine learning', 'explainable AI', 'XAI', 'natural language processing' and 'food delivery services' from Scopus database. This review focused on different ML and DL techniques used in customer sentiment analysis in FDS and selected papers on XAI, DL model and NLP task. A total of 97 papers published from 2001 to 2022 were found and considered for the aforementioned task. Step 2 is described in the 'Results' section.

Step 3: The report phase involves a discussion of the findings, assessment, recommendations and conclusions identified from the research and review papers. This review concludes with the future research direction of increasing the accuracy and explainability of DL models with the help of XAI.

Figure 2.1 shows the various steps involved during literature review. Over 200 scientific papers have been reviewed out of which 97 papers were selected for detailed review on this research topic.

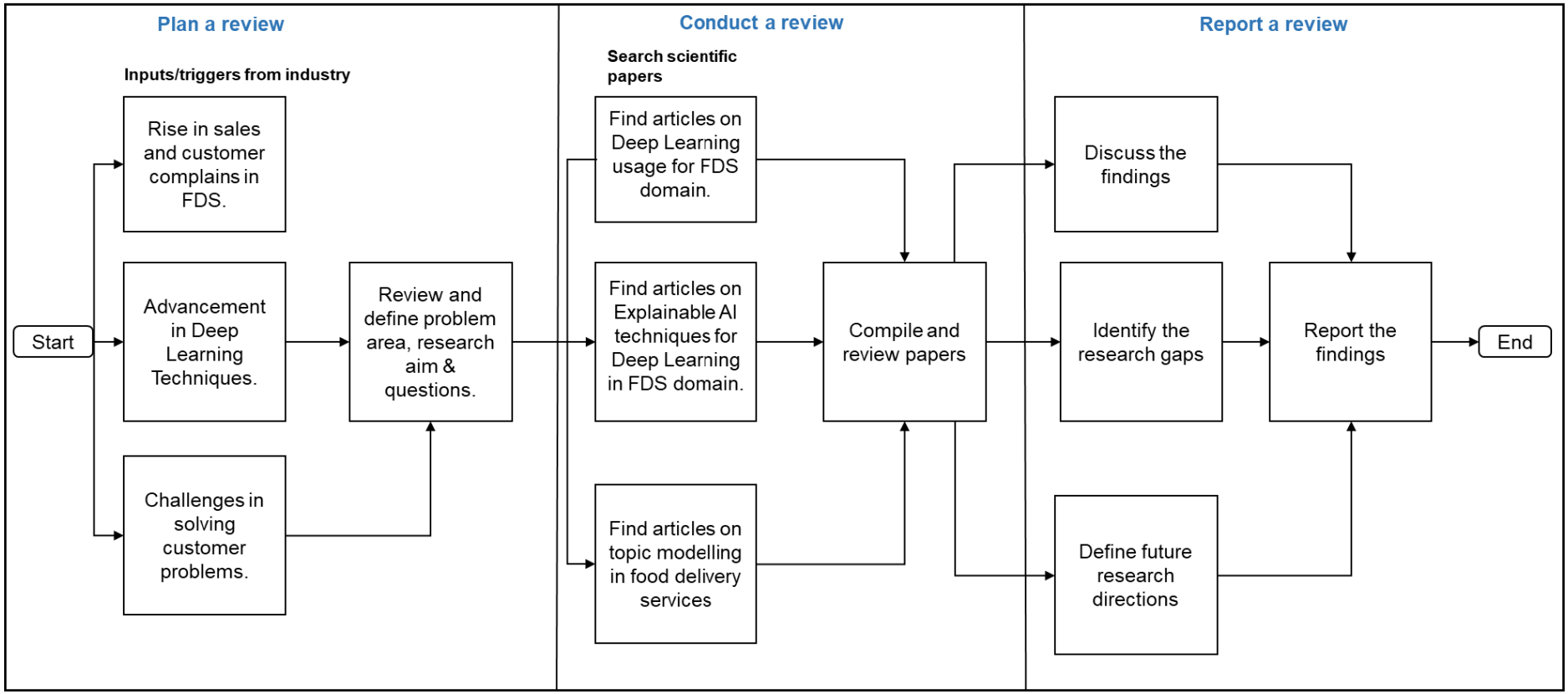


Figure 2.1. Literature review methodology

2.2.1 Aim and Research Questions

The key motivation for this work is as follows. Studies on the sentiment analysis of FDS showed the usage of data mining and ML techniques but lacked focus on DL methods. Additionally, organisations require decision-making models which are justifiable and legitimate. However, no comprehensive study has been conducted to provide insights into the interpretability of published research and the application of state-of-the-art XAI techniques in the FDS domain.

The objectives of this review are to identify the DL techniques applied in the FDS domain for the sentiment analysis of customer reviews, determine the interpretability of published research, identify XAI techniques applied in the FDS domain to bring out the explainability of the models and answer the following questions:

What are the different AI methods used in the sentiment analysis of customer reviews for FDS?

- Is the research on DL technique adequate to identify the negative sentiments of customer reviews?
- What are the challenges in using DL techniques for businesses?
- Can XAI techniques provide explanation and build trust in the DL model?

2.2.2 Search and Selection Process

Table 2.1 describes the keywords (food, deep learning, machine learning, natural language processing, food delivery services, online food delivery and XAI) were used to search the Scopus library. The keyword search criteria were ‘Search within: Article title, Abstract, Keywords’. Only published and peer reviewed papers were considered for further review. After the list of papers from the search results was skimmed, the papers were classified into four categories as shown in Table 2.2.

Table 2.1. Search queries and results showing the number of papers.

No.	Search Query	No. of papers
1	'Sentiment Analysis of customer reviews' AND 'food'	47
2	'Sentiment Analysis of customer reviews' AND 'food' AND 'deep learning'	5
3	'Sentiment Analysis of customer reviews' AND 'food' AND 'machine learning'	18
4	'XAI' AND 'deep learning' AND 'natural language processing'	6
5	'Sentiment Analysis' AND ' Food Delivery Services'	7
6	' Sentiment Analysis' AND ' Online Food Delivery'	8
7	'XAI' AND 'Food'	5

Table 2.2. Literature classification.

Paper Classification	Machine Learning	Deep Learning	Explainable AI Methods	Other Methods	Total
Duplicate papers	18	6	1	15	40
Non-relevant to FDS	9	1	10	10	30
General FDS paper	8	4	0	13	25
Total	35	11	11	38	95

Among the 95 papers, 40 were classified as duplicate from different search queries and hence were excluded from further review. Additionally, 25 papers were found to be

generally related to the FDS domain, and a few were referred to establish context as necessary. These papers were searched and retrieved separately from the University of Technology Sydney library, internet and organisation websites.

2.3 Previous work on FDS using sentiment analysis

Sentiment analysis can be characterised into two primary classifications: traditional approach lexicon-based and machine learning techniques with ML/DL methods see Figure 2.2.

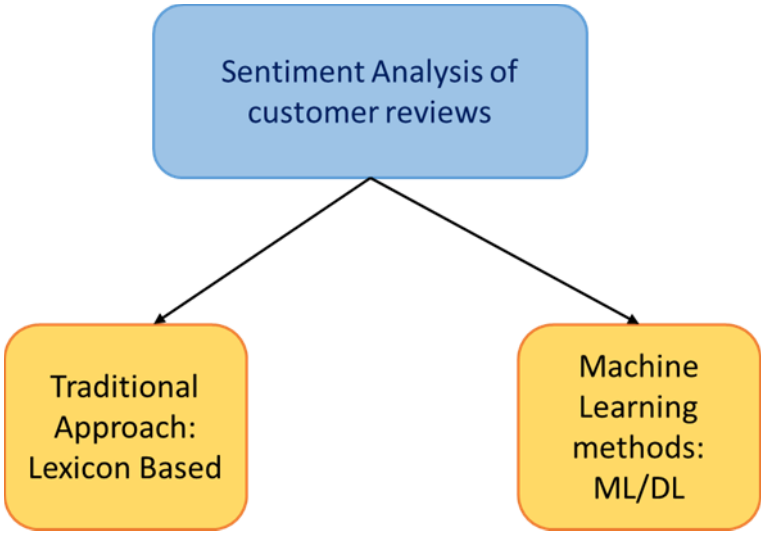


Figure 2.2. Classifications of techniques for Sentiment Analysis.

The below subheadings would describe the lexicon and machine learning based techniques from different review papers.

2.3.1 Traditional approaches on FDS using sentiment analysis

Lexicon-based techniques use a variety of words and their orientation (positive, negative, etc.) to categorise a given text into the correct class (Windasari and Eridani 2017). A lexicon can be used to identify the terms in a document using the bag of words approach. By combining the data and applying a merging method, for example the established average of every class, the phrase's overall sentiment can be predicted.

Most of methods for creating sentiment lexicons rely on dictionaries or corpora. The first method makes use of a lexicon where the words are labelled with their previous polarity.

WordNet (Fellbaum 2017), which connects adaptive synsets, is a nice example of a lexicon in this field. Each WordNet synset (positivity, negativity, and neutrality) is given three numerical emotion ratings by SentiNet (Alshari et al., 2018). The MPQA opinion corpus (Alshari et al. 2018) does this by providing a list of words along with its PO Stagging that are labelled with (positive, negative, and neutral) polarity and (positive, negative, and neutral) to showcase strong or weak intensity. To assess the sentiments of a new document or phrase, the synonyms of names, verbs, adjectives, and adverbs are compared to the seed words that were previously labelled in the lexicon.

The second method looks for sentiment polarity in a domain corpus. There are chances of positive and negative words appearing together in search engines (Ravi and Ravi 2015). The name of the type of the relationship, the relationship's governor, and the dependency of the relationship are all addressed by corpus-based approaches (Xiao and Guo 2015). Even though the systems don't require training datasets, nonetheless they are under pressure to adapt to new data patterns as a result of linguistic change, the expansion of high-dimensional data, the structural and cultural intricacies of short text like the usage of emoticons, tweets and abbreviations (Krouska et al. 2020). Additionally, as each domain has a unique meaning, sentiments formed from one domain could not be appropriate to another (Krishnakumari et al. 2020). sFor example, the term "lightweight" in the context of kitchen appliances may evoke a negative response, whereas the same term in the context of electronics and mobile appliances will elicit a positive response. FDS organisations must use a cross-domain sentiment adaptation ML/DL classifier that is applicable to any domain to solve this problem. When it comes to the difficulty of quickly evaluating these novel data kinds, machine learning algorithms have shown to be better to lexicon-based approaches.

2.3.2 Machine learning approaches on FDS using sentiment analysis

2.3.2.1 Support Vector Machine

It has been demonstrated that support vector machines (SVMs) outperform Naive Bayes (NB) in traditional text categorisation (Joachims 1998). SVM (Ali et al. 2019) is well-known for resolving two-group classification problems in a quick and reliable manner.

The classification is carried out to find the hyperplane between the positive and negative reviews of two classes in the model, which is given in eq. 2.1 as:

$$B = \min_{i=1\ldots m} |w \cdot x + b|. \quad \textbf{(2.1)}$$

Every category has a B_i value for the number of hyper planes, which is denoted by s, for category classification. As a result of this research model's success in determining the largest B_i value is given in eq. 2.2 as:

$$H = \max_{i=1\ldots s} \{h_i \mid B_i\}. \quad \textbf{(2.2)}$$

After standard length-normalizing the document vectors, Pang et al. (2002) used SVM^{light} (Scholkopf 1999) for package 8 for training and testing, with all parameters set to their default values (neglecting to normalise generally hurt performance slightly).

2.3.2.2 Naïve Bayes

Assigning the class $c^* = \arg\max_c P(c \mid d)$ to a given document d is one method of classifying text. We first observe that by applying eq. 2.3 to the Bayes' rule, the following results are obtained:

$$P(c \mid d) = \frac{P(c)P(d|c)}{P(d)}. \quad \textbf{(2.3)}$$

where $P(d)$ has no role in selecting c^*. Using eq. 2.4 and the assumption that f_i 's are conditionally independent given d's class, Naive Bayes evaluates the term $P(d \mid c)$, as follows:

$$P_{\text{NB}}(c \mid d) := \frac{P(c)\left(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)}\right)}{P(d)}. \quad \textbf{(2.4)}$$

Despite its simplicity and the fact that its conditional independence assumption manifestly fails in real-world situations, Naive Bayes-based text categorisation performs admirably (Lewis 1998), and research (Domingos and Pazzani 1997) shows that Naive

Bayes is best for some problem classes with highly dependent features. SVM and Maximum Entropy, on the other hand produces better results.

2.3.2.3 Maximum Entropy

Maximum entropy classification (ME) has been shown to be useful in a variety of natural language processing applications (Berger et al. 1996). According to research (Nigam et al.), it occasionally performs better than Naive Bayes in classifying standard text. Its $P(c \mid d$ estimate has the following exponential form is given by eq. 2.5 as:

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right) \quad \textbf{(2.5)}$$

where $Z(d)$ is a normalization function. $F_{i,c}$ is a feature and class function for combined feature f_i and class c, defined in eq. 2.6 as follows:

$$F_{i,c}(d,c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases}. \quad \textbf{(2.6)}$$

For instance, if the bigram "still hate" appears and it is assumed that the page is filled of hatred, a certain feature or class function could activate. Importantly, MaxEnt does not make any assumptions about feature connections, in contrast to Naive Bayes, and could thus perform better when conditional independence assumptions are broken.

According to the P_{ME} definition , $\lambda_{i,c}$'s feature-weight parameters; a large $\lambda_{i,c}$ indicates that f_i is a strong indicator for class c. The parameter values are selected with the predicted values of the feature/class functions in mind in order to maximise the induced distribution's entropy (thus the classifier's name).

2.3.2.4 Deep learning techniques

Another study (Luo and Xu 2019) indicated the success of DL models which comprise hundreds of layers and parameters and outperform traditional ML algorithms in sentiment classification and review rating prediction. Some challenges arise with DL usage, such as the requirement for large data, heavy computing and training models. Nevertheless, in to-day's world, these challenges are no longer an issue because of the availability of high-performance computing facilities.

- ***Recurrent Neural Network (RNN)***

RNN is a class of neural networks which works well with a sequence of data input (Lopez and Kalita 2017). NLP tasks, such as sentiment analysis, can be easily solved by RNN. Different from traditional neural networks, RNN can remember the previous computation of information and can apply it to the next sequence of inputs.

According to some researchers (Luo and Xu 2021; Tian, Lu, and McIntosh 2021; Luo et al. 2020; Zahoor et al. 2020; Hegde et al. 2018; Shaeeali et al. 2020; Drus and Khalid 2019; Pang et al. 2002; Lopez and Kalita 2017), DL algorithms (bidirectional long short-term memory (Bi-LSTM) and simple embedding and average pooling) performs better than traditional ML algorithms in sentiment classification and review rating prediction. They proposed the use of DL technique during the COVID-19 pandemic to help customers in making safe dining decisions. The review data were obtained using a web scraper from Yelp restaurants located in the top 10 cities by population in the United States and were pre-processed by tokenisation and stopword removal (Tian, Lu, and McIntosh 2021; Luo et al. 2020; Zahoor et al. 2020; Hegde et al. 2018; Shaeeali et al. 2020; Drus and Khalid 2019; Pang et al. 2002; Lopez and Kalita 2017; Suciati and Budi 2020; Molnar 2020). Term frequency-inverse document frequency was used to identify the key features from the reviews and place them into meaningful categories. The results showed that the basic embedding and average pooling perform well in online review prediction tasks, whereas the bidirectional LSTM method is successful in generating subtopics and sentiment prediction. Luo and Xu (2021) suggested that RNN models require a high level of supervision and that future works should focus on the bidirectional RNN model.

A systematic review on sentiment analysis in social media conducted by (Drus and Khalid 2019) revealed that RNN has a longer computational time than other DL models (convolution neural network, CNN). Common DL models such as RNN, LSTM and CNN have been individually tested in different datasets; however, their comparative analysis is lacking. Tian et al. (2021) highlighted that DL models such as RNN is efficient in handling a large volume of complex data but is often criticised for being a black-box model. Using DL models in FDS domain (Adak et al. 2022; Luo and Xu 2021) have

interpretability issues, XAI techniques SHAP and LIME was implemented to overcome the problem of interpretability of the black box DL models. There are several XAI techniques re-searched so far in other domains, but best to our knowledge none is applied into FDS domain. Further work must be conducted for the comparative analysis of DL models in performing sentiment analysis in the FDS domain.

- ***Convolution Neural Network (CNN)***

CNN is widely popular because it can be used in image datasets by extracting the significant features of the image while the 'convolutional' filter (i.e., kernel) moves through the image (Ajit et al. 2020). CNN could also be used in text with 1D input data (Johnson and Zhang 2014). While the filter moves in the text area, the local information of texts is stored, and important features are extracted. Hence, CNN can be effectively used for text classification. Kim (2014) found that CNN models outperformed previous approaches for several classification tasks. With the slight tuning of the hyper-parameters, one-layered CNN performs remarkably well. Moreover, unsupervised pre-training of word vectors plays a key role in DL for NLP. Bhuiyan et al. (2020) found that attention-based CNN model had the highest accuracy of 98.5% compared with that of baseline CNN at 96.34% and LSTM at 97.23%. They proposed to work on the usage of bidirectional encoder in the FDS do-main because it produces the best results with extremely long training time compared with CNN. Hung (2020) indicated that the hybrid model of CNN with LTSM is more accurate than CNN or LTSM. The accuracy of the hybrid model is 83.45%, whereas that of individual CNN and LSTM is 82.76% and 82.54%, respectively. Muhammad et al. (2020) compared the performance of various ML algorithms such as SVM, logistic regression, random forest and NB and found that the CNN model outperformed all ML algorithms. Therefore, CNN can be used in text mining tasks with high accuracy and could be applied for customer sentiment analysis on FDS.

According to the literature, hybrid DL models should be tried to attain accuracy in performing sentiment analysis. Additional research must be conducted to improve the interpretability of the black box models of DL algorithms.

2.3.3 Explainable AI techniques

The success of DL models which comprise hundreds of layers and parameters considered as a black box (Arrieta et al. 2020). Organisations need models capable of making decisions which are justifiable and legitimate. A common perception is that if the model only targets accuracy and performance, then the system would become opaque. However, understanding the model features would enable the improvement of its deficiencies. According to Singh et al. (2020), DL is significant in medical diagnostic tasks and outperformed human experts. However, due to the black-box nature of the algorithm, it is not being used across the industry. The study (Wolanin et al. 2020) signified the importance of ML and DL in the context for forecasting crop yields (different domains) but added that these algorithms lack transparency and interpretability. The black-box nature of DL restricts its usage across the industries because it lacks trust and explainability.

Interpretability is the degree to which a human can comprehend the reason for the model's outcome (Molnar 2020). Deep neural networks lack interpretability, and the model features that drive the outcome are difficult to understand (Guidotti et al. 2018; Kenny et al. 2021; Liz et al. 2021; Lorente et al. 2021; Moradi and Samwald 2021; Samek et al. 2021). XAI or interpretable machine learning IML programs strive to create models that are explainable while keeping a high level of accuracy. Study (Schoenborn and Althoff 2019) indicates that the need for explainable AI has increased rapidly due to the increase in usage of DL and recent legal restrictions. The goal is to bring people to trust AI which can be achieved through explainable AI. In implementing DL models, we need to provide explainability on how the model predicts its outcome so that industries and organisations can build trust to apply the black-box model. A possible scenario is that a DL model has extremely high accuracy for wrong reasons and organisations cannot trust any model without knowing which feature or dimension served as the basis of the prediction. According to Mathews (2019), black boxes should not be employed in critical systems such as the medical profession or malware detection since incorrect judgments might have serious effects. Most research in FDS achieved accuracy with non-interpretable models. Table 2.3 shows the recent papers on sentiment analysis in FDS with model interpretability, results and future directions.

Table 2.3. Interpretability of methods used for sentiment analysis in FDS.

Paper	Algorithm	ML/ DL	Year	Is Method Interpretable	Refs	Results
Comparative study of deep learning models for analysing online restaurant reviews in the era of the COVID-19 pandemic	Bidirectional LSTM and Simple Embedding + Average Pooling	DL	2021	No	(Luo and Xu 2021)	Study finds DL models perform better than ML models. Bi-LSTM model (92%) scores over Simple Embedding +Average Pooling model (90%), GBDT (88.9%), Random Forest (86.6%). Future work recommended to be done to explain the DL black boxes as they are non-interpretable. Also research was carried out in limited location dataset, hence focuses on the importance of implementing it in larger scale in terms of location.

Integrating Sentiment Analysis in Recommender Systems	LSTM,CNN,LSTM-LSTM	DL	2020	No	(Hung 2020)	Study finds hybrid model of CNN-LSTM model (83.45%) scores more than LSTM (82.54%) and CNN (82.76%). Future work recommended on learning content features to matrix factorization latent factors.
Aspect-based sentiment analysis and emotion detection for code-mixed review	Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM)	DL	2020	No	(Suciati and Budi 2020)	Study finds Random Forest dominates with 88.4% and 89.54% F1 scores with CC method for food aspect, and Label Powerset for price. For service and ambience aspects, Extra Tree Classifier leads with 92.65% and 87.1% with Label Powerset and Classifier Chain methods, respectively whereas in deep learning comparison, GRU and BiLSTM obtained similar F1- score for food aspect, 88.16%.

						On price aspect, GRU leads with 83.01%. However, for service and ambience, BiLSTM achieved higher F1-score, 89.03% and 84.78%. Drawback of the study was it used unbalanced dataset and hence should be considered for oversampling or undersampling.
An Attention Based Approach for Sentiment Analysis of Food Review Dataset	CNN, LSTM and CNN + Attention	DL	2020	No	(Bhuiyan et al. 2020)	Study finds hybrid model of CNN + Attention to acuquire 94% as compared to LSTM 93% and CNN 91%. Future work towards usage of BERT architecture is recommended provided large dataset is gathered.
Sentiment analysis and classification of restaurant reviews using machine learning	Naïve Bayes Classifier, Logistic regression,	ML	2020	No	(Zahoor, Bawany,	Study find Random Forest to acquire 95% accuracy on the dataset when compared with ML models. Future work towards deep

	Support Vector Machine (SVM), and Random Forest				and Hamid 2020)	learning and neural networks is recommended.
'How was your meal?' Examining customer experience using Google maps reviews	Logistic regression	ML	2020	No	(Mathayomchan and Taecharungroj 2020)	Study used Vader framework to measure sentiment of four key restaurant attributes: food, service, atmosphere and value. Study points to the platform bias as only google map users were able to put review. Also, data was collected from larger cities and need to go smaller city restaurants. Future study recommended in other tools.
Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia	Logistic regression, Decision tree	ML	2019	No	(Suciati and Budi 2019)	Study used ML models for comparison and found Logistic regression scored 81.76% accuracy for food and 77.29 % for ambience and Decision Tree scored 78.71% for price and 85.07% for service aspects. Future work

						recommended on Deep learning methods to achieve higher accuracy.
Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Naïve Bayes	Multinomial naïve Bayes	ML	2019	Yes	(Sharif, Hoque, and Hossain 2019)	Study used Multinomial naïve Bayes model to get 80.48% accuracy and future work is recommended on usage of powerful algorithms. Getting large dataset is one of the primary concern added in this study.

Table 2.3 shows that 45% of the papers used a model built on DL and 55% used a model built on ML. Many studies (Bhuiyan et al. 2020; Zahoor et al. 2020; Suciati and Budi 2019; Sharif, et al. 2019) suggested the need for usage of DL methods in their future work. Another study (Luo and Xu 2021) applied DL methods in the study and found DL methods were working better than ML methods. However, they emphasised the need for explainability of the DL models as DL models are black box in nature and has no interpretability. The key fact is that 77% of the models are non-interpretable in nature; hence, organisations can argue for the explainability and trust in the system. No study has been conducted on XAI with DL on NLP for sentiment analysis across the FDS industry, which represents a scope for future research. Many XAI methods can be applied to DL models to increase the explainability component and ensure high accuracy. The most popular two XAI methods are the following.

2.3.3.1 Local Interpretable Model-Agnostic Explanations (LIME)

Shankaranarayana and Runje (2019) proposed a method called LIME (Ribeiro et al. 2016). LIME is one an XAI technique that generates single-instance level explanation by artificially generating a dataset around the instance (by randomly sampling and using perturbations) and then training a local linear interpretable model. For sentiment analysis, organisations need to understand the words or features which contribute greatly in predicting the reviews to be negative, neutral or positive. Given the previous application of LIME in other domains (Mathews 2019; Utkin et al. 2020), it can be used in DL models to analyse customer reviews in the FDS domain. No research has been published on sentiment analysis in FDS and DL along with LIME interpretability.

2.3.3.2 Shapley Additive Explanation (SHAP)

SHAP is based on the principle of adding the SHAP value as a contribution to all the variables of a data point to derive the final outcome (Lundberg and Lee 2017). This technique functions in the same way as any team sport, such as cricket or football. Once a cricket match is completed, post-match analysis can be performed using a SHAP-based algorithm. For any outcome such as win, lose or draw, contributions from all 11 players can be used to evaluate the SHAP value for each player. Internally, SHAP uses Kernal SHAP method from, which computes the weight as a contribution for all the features of

the black box (Kim et al. 2016). SHAP is built to enhance the features of LIME. Different from that in LIME, a local linear module is not built in SHAP. Instead, some functions are used to calculate the shapely value. In sentiment analysis, the SHAP algorithm can be used to determine the contributions of each word towards positive and negative sentiment. However, no research has been conducted on sentiment analysis in the FDS domain and DL along with SHAP interpretability.

2.3.3.3 Comparison of LIME and SHAP

The major difference between LIME and SHAP is that the LIME value is evaluated by removing the variables or features to obtain an outcome, and the SHAP value is the contribution of all the variables or features to make a prediction (Psychoula et al. 2021). Owing to this nature, LIME is much faster than SHAP because the latter considers all the possible combinations of the variables with contributions to create the outcome.

2.3.4 Topic Categorization

Text mining to the data is the prerequisite in order to analyse topic categorization of the customer reviews (Westerlund et al. 2019). Text mining is a technique for analysing and classifying text data by calculating statistical estimates based on the frequency and ratios of words that appear in the text (Kim and Kang 2018). It attempts to find useful models, trends, patterns, or rules from unstructured textual data, with the goal of extracting meaningful information from a large number of documents (He et al. 2017). There are, however, a variety of text mining methods and approaches. Probabilistic topic modelling, an unsupervised machine learning method, has grown in popularity as a tool for text mining in social science research over the last ten years (Schmiedel et al. 2019). The basic concept of topic modelling is that each document can address a variety of topics that are unknown in advance (Bittermann and Fischer 2018). Because text documents are composed of words, and a topic spoken in multiple documents can be expressed by a combination of strongly related words, topic modelling aids in inferring hidden topics in text documents (Jeong et al. 2019). Because each document is assumed to address each topic to varying degrees (0-100 percent), each document can belong to multiple topics (Jeong et al. 2019; Bittermann and Fischer 2018).

Latent Dirichlet Allocation (LDA) is by far the most common topic modelling algorithm (Calheiros et al. 2017; Kim and Kang 2018). LDA is considered to outperform other topic modelling algorithms when dealing with large-scale documents and interpreting identified latent topics (Jeong et al. 2019). It applies a relational approach to meaning in that word co-occurrences are crucial in building their meaning as well as the meaning of subjects (Schmiedel et al. 2019).

When given a probabilistic distribution of the document's topics and a probability distribution of the words that make up each topic, LDA stochastically chooses the document's topics and repeatedly samples the words in the selected topics (Kim and Kang 2018). According to Calheiros et al. (2017), LDA not only allows determining the likelihood of a selected review belonging to each topic and grouping reviews based on their proximity to each considered term, but it also aids in identifying which topics are attracting more attention.

In their article, Brandt et al. (2017) used Latent Dirichlet Allocation on geo-tagged social messages to add value to the smart tourism ecosystem. They looked at the spatial dimensions of LDA topics in and around San Francisco. To figure out what the person is talking about, they extracted topics based on the location or the location tagged on the tweet. They cleaned the data, created a corpus, and used topic models in the R programming language to extract the top 30 topics. They attempted to comprehend urban dynamics using the insights generated by the topic models. LDA could assist them in determining how a specific event will elicit excitement in the crowd in the area where the event is scheduled to take place in the near future or after it has taken place. These insights are especially useful for businesses looking to add value to their customers and increase citizen engagement.

A flaw in using Latent Dirichlet Allocation with tweets has been identified in their study (Abdelwahab et al. 2014; Brandt et al. 2017). According to them, LDA assumes that the text contains a number of different topics that are grouped together to form the document. For LDA, the text in a tweet is insufficient. To overcome this limitation, Abdelwahab et al. (2014) grouped tweets from the same country together so that the LDA could process them.

2.4 Strength and limitations of models

This study showed that the performance of ML models (Naïve Bayes, maximum entropy classification and SVM) on sentiment classification is not as good as that of traditional topic-based categorisation (Pang et al. 2002). Customer reviews can be negative without having any negative word in the sentence. Additionally, lexicon-based approaches can achieve higher accuracy than ML models but are challenging to implement in sentiment analysis in languages other than in English (Krishnakumari et al. 2020). Domain adaptation is another aspect which must be considered in building models because the same words can have different meanings in another domain. The mentioned challenges may be solved by using DL algorithms where the model trains itself from a large chunk of data from the same domain.

DL methods such as RNN, CNN, and LSTM showed good performance. However, further experiment and research must be conducted on hybrid approaches where multiple models and techniques are combined to enhance the sentiment classification accuracy (Dang et al. 2020). Although neural networks provide high prediction accuracy (Kim 2014), they lack explainability. Owing to the opaqueness of the DL techniques, businesses are reluctant to use black-box models and prefer to verify and check how the models are predicting accurate results. XAI techniques such as SHAP and LIME can support DL techniques in explaining how the model is determining the correct customer sentiment of a review. LIME and SHAP results can be compared with those from DL techniques.

By performing sentiment analysis using the DL/ML methods on customer reviews, FDS organisations can use the data to analyse customer complaints and work towards improving customer satisfaction. The output customer review data from DL/ML model is labelled as negative and positive sentiment. The ML/DL model is verified using the XAI technique against its computing logic. As topic modelling can group related topics, the negative sentiments can be grouped into different classes (delivery time, customer service, food quality and cost) as shown in Table 4. FDS organisations can use this information to understand which particular group class is getting more problem. Different problem categories may be sent to the respective team. If the negative sentiments are due

to an increase in delivery time, then organisations may need to solve their supply-chain-related problems. FDS organisations may also look into logistics issues by determining the number of vehicles and delivery boys needed when delivering to far-off destinations.

In case of large orders in restaurants, delivery time sometimes increases due to larger wait time. The higher delivery time data may be further grouped upon location to check if the problem is happening for some locations or all locations. If the negative feedback comes under customer service category, then the service level must be paid attention. With food delivery, there is always a risk of poor packaging or spillage and hence food quality issues must be resolved at the respective restaurants, and organisations can keep an eye on the restaurants which are contributing to negative reviews due to food quality. Complains on the cost of the food item can be resolved by the restaurant and the organisation by reducing the cost or lowering the profit margin. Several other complaint groups can be considered by the FDS organisation to solve their customer feedback complaints. Topic categorisation on positive sentiments can also be used to reward staffs or restaurants. FDS organisations may think of more meaningful topic groups based on their business requirement. Although topic modelling has performed significantly well in topic categorisation, but there is a need to compare these techniques on the FDS domain.

Although customer feedback or reviews are easily obtained from blog posts, comments, reviews or tweets, the data can be of a very large volume. DL models have always shown good performance with a large volume of data. Thus, new DL or hybrid models should be tested to obtain the best accuracy. The negative sentiments can be categorised into various complaint groups using topic modelling. For the DL models, explainability must be reduced to achieve high accuracy; however, XAI can support the explainability part of the model. Several research papers have presented the usage of ML or DL techniques for sentiment analysis in customer reviews; however, no study has been conducted on XAI with DL in the FDS domain. With the surge in FDS usage due to COVID-19 lockdowns, the solution (see Figure 1.2) can definitely help the food industry to quickly adapt to customer requirements and preferences.

2.5 Current research issues in food delivery services

Although customer feedback or reviews are easily obtained from blog posts, comments, reviews or tweets, the data can be of a very large volume. DL models have always shown good performance with a large volume of data. According to literature review, no evidence is found on the application of XAI techniques on DL models in the FDS industry to analyse customer reviews. In the past, work on sentiment analysis of customer reviews of FDS is done using traditional based techniques like lexicon based or machine learning classifiers. Since both lexicon and machine learning techniques face issues such as domain adaptation, it is worth to look at DL models to solve the problem. Also, it is important to check if the DL models are able to predict the sentiments accurately on customer reviews from FDS domain. If the accuracy is accepted, it is difficult for FDS organisations to trust the black box DL models as they are not interpretable. There are no papers available on implementation of XAI techniques to unbox the DL models in FDS domain to verify the prediction logic. Once the problem is identified in the form of negative sentiments, FDS organisation need to know on what aspects are causing the customer complaints. There should be some mechanism to go through the full set of negative customer reviews (complaints) and pick the common issues causing the problem. Problems can be due to customer service or food taste or it can relate to delivery time. Similarly, with positive reviews, FDS organisations can look into the things which are going in the right direction for their business and reward the restaurants or their staff.

2.6 Summary

Compared with ML techniques, DL is more accurate in predicting customer sentiment analysis. Given that deep neural networks are black-box in nature, DL models need support from XAI techniques, such as LIME or SHAP, to explain the features on which algorithms are computed to ensure high accuracy and explainability and earn the trust of businesses. The combination of DL with XAI on FDS would help in understanding the customer sentiments about food and service quality worldwide and subsequently improving customer satisfaction. Furthermore, topic modelling can be conducted on customer reviews to categorise them in meaningful groups. According to the volume of

complaints in each group, organisations can prioritise their action and send it to the right channel for a solution.

Results showed that DL techniques (CNN, LTSM and Bi-LTSM) have great accuracy but lack explainability; their interpretability can be improved with XAI implementation. Domain adaptation by the models is a key aspect in sentiment analysis. In consideration of the increase in sales and competition across this domain, additional research work is required on sentiment analysis in the FDS domain using DL techniques with XAI. Thus, the following research directions are recommended:

- Further research on the sentiment analysis of customer reviews using DL techniques such as CNN, LTSM and Bi-LTSM and comparison of the results;
- Usage of XAI techniques such as LIME or SHAP to explain and build trust in the DL models from the previous step;
- Classification of negative sentiments into various topic categories using topic modelling techniques to address supply chain issues and improve customer satisfaction; and classification of the positive sentiments into various topic categories using topic modelling technique to appreciate or reward employees.

3 MATERIALS AND RESEARCH METHODOLOGY

3.1 Introduction

This chapter illustrates several methods applied in this study. Overall methodology, implementation of the detailed methodology and performance evaluation are described. The data acquisition as how it was scraped from website along with key attributes and data cleansing process is described. The deep learning algorithms, XAI methods and topic categorization technique is introduced to analyse sentiments from customer reviews, perform validation on the black box models and pick topics from the customer complains to resolve supply chain issues. LSTM, Bi-LSTM and Bi-GRU-LSTM-CNN model was developed for performing sentiment analysis on customer reviews. XAI methods such as SHAP and LIME was implemented on DL models to measure the accuracy by validating the features on which the outcome was predicted by DL models. LDA technique was built to pick topics on the customer complaints to identify key areas for improvement. All the objectives mentioned in the research in introduction section are addressed by implementation of the detailed methodology.

3.2 Data Acquisition

3.2.1 Data scraping using ParseHub

Productreview.com.au is an Australian website that gathers consumer feedback on a variety of products and services. Overall, 13,621 customer reviews were collected from various FDS companies, such as Uber Eats, Menulog, Youfoodz, Deliveroo, My Muscle Chef and Macros, from the ProductReview website via web scraping. Figure 3.1 of the product review website for ubereats:

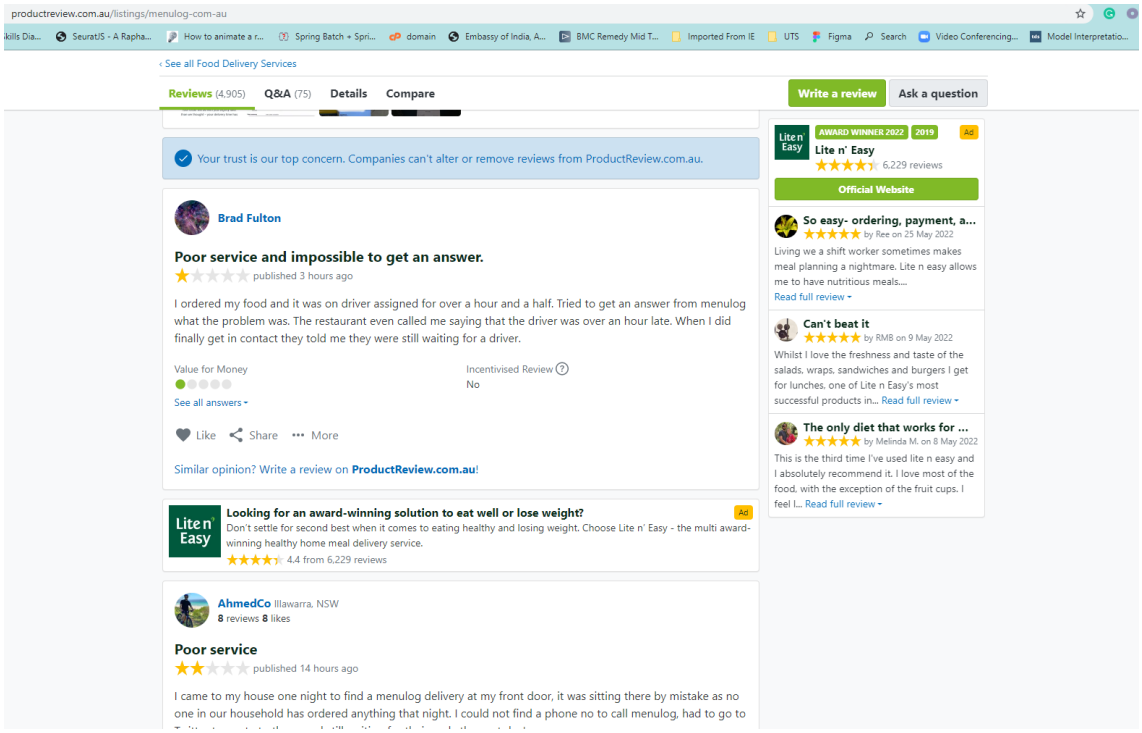


Figure 3.1. ProductReview website for Menulog (www.productreview.com.au)

The customer review data for each customer was scrapped using ParseHub tool. Similar activity was done for all the FDS organisations such as such as Uber Eats, Menulog, Youfoodz, Deliveroo. Figure 3.2 shows the word cloud made by the FDS customer review data.

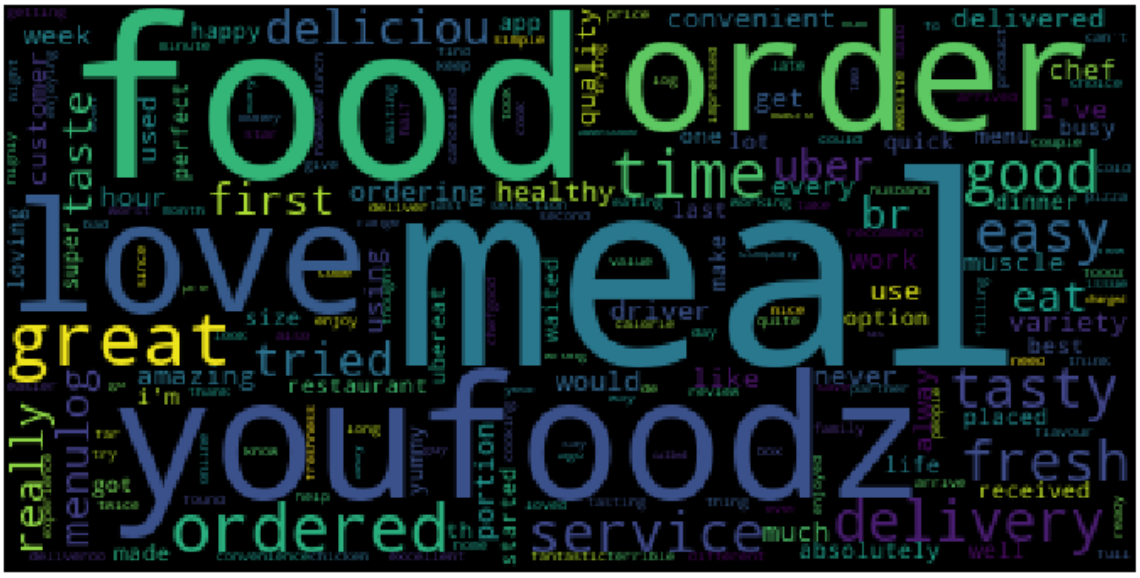


Figure 3.2. Wordcloud of customer reviews from productreview site.

3.2.2 Identify Data Attributes

The below dataset example Table 3.1 shows the various attributes present along with customer reviews. Review comment and star ratings attributes were used to train the DL models.

Table 3.1. Different attributes of the dataset from ProductReview.

Username	Review Topic	Location	Star ratings	Date	Review comment
Mark A	Rubbish!	Sydney, NSW	1 star	3rd Feb, 2021	Hands down the worst delivery service, always slow, food or drinks missing, food cold... Won't be using again, after waiting an hour for food I don't want a voucher just to reorder.
Lucy				30th Jan, 2021	Happy I got an actual refund instead of credit for once, but unhappy that

	Refunding is a joke	Sydney, NSW	2 star		they refused to refund me for ruined food caused by thoughtless and inappropriate packaging. Just use any other delivery app.
Les	When it actually works not a bad site to order food	Greater Melbourne	3 star	13th Sep, 2021	This week all the Melbourne restaurants randomly drop off the site, and take some time to return. If you are unfortunate enough to have an order confirmed, it sits there unmoving until you call them and cancel it, so obviously it has no real recovery mechanism built into the software. Customer service people do their best but often they can only suggest cancelling the order and tying again later. Amateurish at best.
Lily	Good Service	Sydney, NSW	4 star	13 Feb, 2021	Food was on time and hot, the ordering process was slightly confusing but other than that, it was great, good customer service and accurate tracking time! Definitely would use again.
Russell G.	First Time User	Sydney, NSW	5 star	20th Jan, 2021	I provided a wrong address by accident. Driver called me up, advised how far away he was, met me at the door. Food is warm and well packaged - Happy with the Menulog service - will use again.

3.2.3 Data Splitting

The productreview scrapped dataset was grouped into binary sentiment tasks: positive and negative classes. The positive class was labelled from rating 4 and above, and the negative class was labelled from rating 2 or below. The dataset was then divided into 8,995 positive reviews and 4,626 negative reviews as shown in Figure 3.3. Rating 3 was not placed in any of the classes.

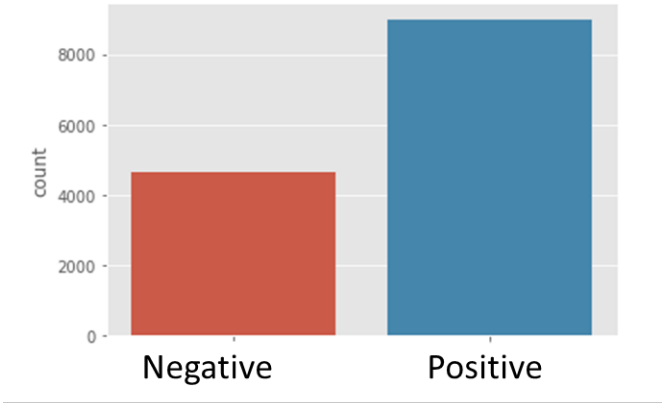


Figure 3.3. Negative and positive sentiment count.

3.2.4 Data cleansing

The labelled customer review data were cleaned by reducing the noise and normalising each word to the lowercase. Further punctuations, such as question marks, commas, colons, hash signs and website URLs, were removed to reduce the noise of the data. Some review data sequences were truncated or padded to have fixed length for making all the sequence data in standard length. For the training data, one of the requirements for LSTM models is to have a fixed length for input sentence length of the review data. We set the customer review data length to 100.

3.3 RNN Architecture

A sequence of data input works well with a RNN (Lopez and Kalita 2017). In traditional neural networks, all the input variables are independent of the output variable. Some of the NLP problem examples, such as predicting if the sentence is positive or negative,

spam classifier or time-series data, stock forecasting or sales forecasting, can be solved by RNN (Yin et al. 2017). Bag of words, term frequency-inverse document frequency and Word2VEC are used for text preprocessing where they convert text into vectors to solve NLP problems in machine learning. The issue with these algorithms is that they discard the sequence information in the sentence, thereby resulting in lower accuracy. The RNN is named after the fact that it performs the same task for each element of the sequence, with the output being dependent on previous computations. RNNs are supposed to have memory that stores information from previous steps. However, they can only look back a few steps in practise (Thikshaja and Paul 2018). Figure 3.4 shows a typical RNN architecture with respect to time-series data.

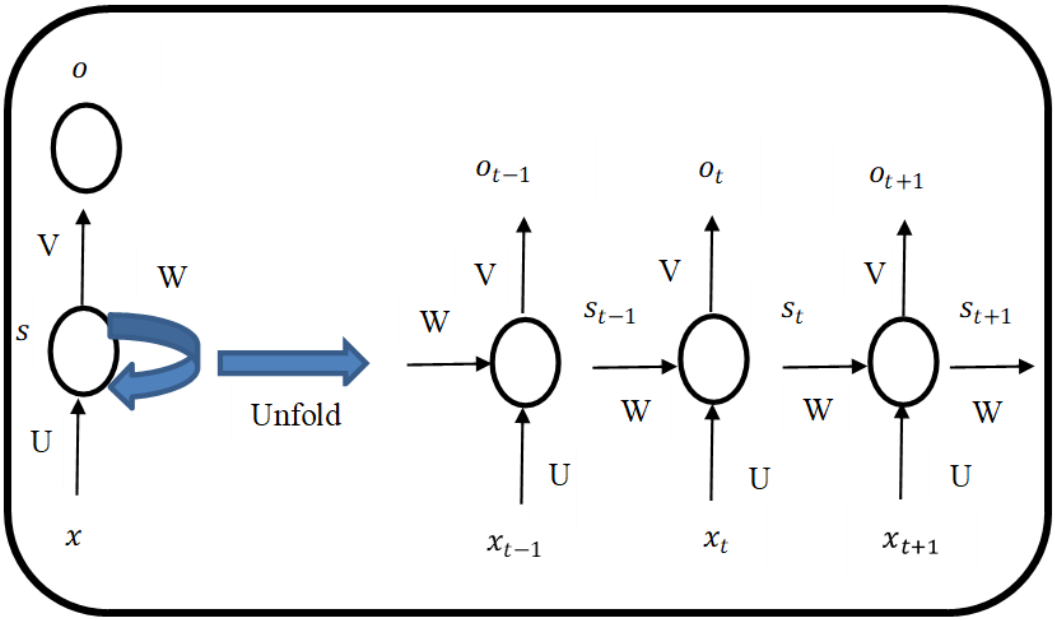


Figure 3.4. Showing RNN architecture.

Assuming we have a sentence of 5 words, then the above Figure 3.4 will have 5 layers, with one layer for each word. In Figure 3.4, x_t is the input, s_t is the hidden state, and o_t is the output step at time step t. The input at time step t is $s_t = f\ (Ux_t + Ws_{t-1})$. The function f is nonlinearity, such as Relu or tanh and s_{t-1}, which is required to initialised to all zeros in calculating the first state.

In a prior research on tweet detection (Cambray and Podsadowski 2019), four deep recurrent architectures based on LSTMs and GRUs were utilised. The model's complexity was boosted by adding convolutional layers. Upadhyay et al. (2022) created a Bi-LSTM model to increase the precision of the sentiment analysis of IMDB reviews using an ensemble of CNN and bidirectional LSTM. The black box models lacked transparency even though they were more precise. In this work, the model was constructed using a Bi-LSTM, GRU, CNN, single LSTM, and single Bi-LSTM with some extra layers. The black box models were tested through the explainability approaches to assess how they performed.

3.4 Deep learning techniques

3.4.1 LSTM and Bi-LSTM

LSTM is a gated RNN, and Bi-LSTM is an extension of the model. LSTM models can learn long dependencies from the previous states as compared to the traditional RNN model (Schmidhuber and Hochreiter 1997). Bi-LSTM model is an extension of the LSTM model, where it trains the input data twice through forward and backward directions. Figure 3.5 shows a typical architecture of the LSTM model.

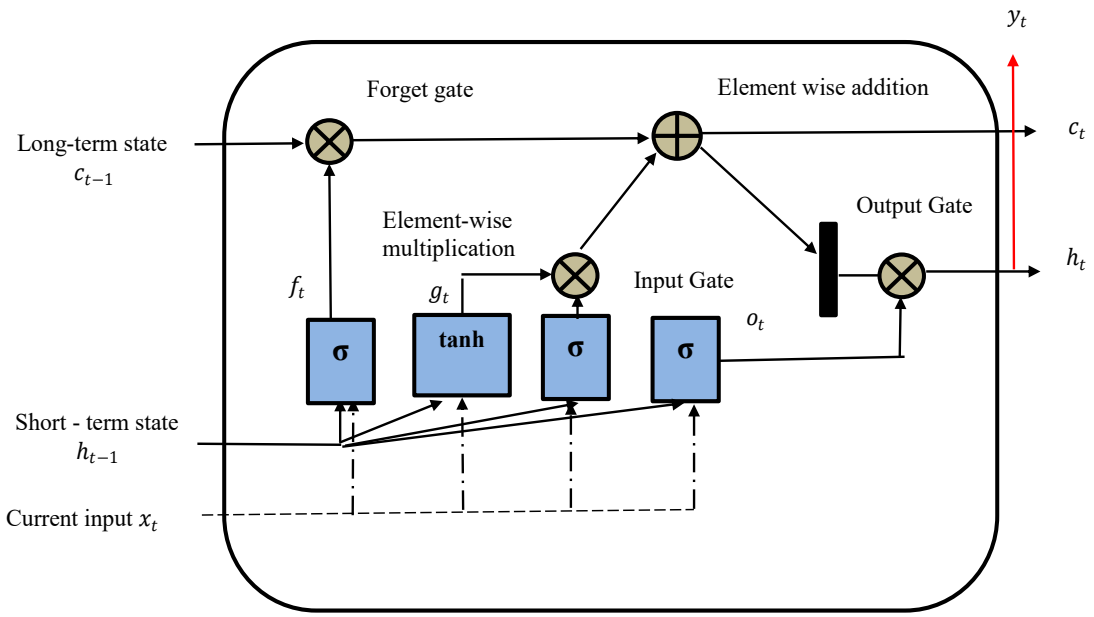


Figure 3.5. LSTM architecture

3.4.2 Bidirectional GRU

GRU, which was introduced in 2014, is similar to LSTM without output gate. GRU has update and reset gates that help in combining new inputs with the previous ones (Cho et al. 2014). The update gate decides how much previous memory is needed to be saved. In LSTM, the cell state and hidden state were known as short-term memory, whereas only one state, that is, hidden state, was found in GRU. GRUs have shown better performance on smaller to medium quantity datasets.

3.5 XAI Techniques

3.5.1 SHAP

SHAP is a game theoretic means to explain any ML models. It explains how to predict an instance x by computing each feature's contribution to the prediction (Ancona, Öztireli, and Gross 2019). Shapley values are perturbation-based methods, where no hyperparameters are required, except for the baseline. The Shapley value eq. 3.1 is calculated as follows:

$$Ri = \sum_{S \subseteq P \setminus \{i\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} \left[\hat{f}(S \cup \{i\}) - \hat{f}(S)\right], \quad \textbf{(3.1)}$$

where P represents a set of N players, and $\hat{f}$ maps each subset of S ⊆ P of players to real numbers. The result $\hat{f}$ (P) of the game is represented by the contributions of all players. The Shapley value for player i can be described as the average marginal contribution of player I to all possible combination S that can be formed without it.

With $\hat{f}$ as the set function, the above equation can be implemented for neural network function f. We replace $\hat{f}$ (S) with $\hat{f}$ (x_s), where x_s indicates the original input vector x with all features not present in S are replaced by the baseline value.

3.5.2 LIME

Lime is a model-agnostic and concrete implementation of local surrogate models. LIME focuses on training local surrogate models rather than global surrogate models to explain individual predictions. Lime tweaks the feature value of a single data sample and checks for the change in the output. Lime generates new texts by removing words randomly from the original text.

LIME generates a collection of scores, defined as E, from a text sequence T and a text classifier C, where the elements indicate the relevance r(t) ∈ [−1, 1] of the word tokens t ∈ T in relation to a specified class c of interest (Di Cicco et al. 2019). Tokens in T that move C's prediction towards c receive a positive score from LIME, whereas tokens in T that move C's prediction towards to any other class c ′ c receive a negative score from LIME. The tokens in T are assigned positive score by LIME that drive the prediction of C in the direction of c and negative score to tokens that push to any other class c ′ ≠ c.

3.6 LDA

One of the most often used topic modelling techniques is LDA (Jelodar et al. 2019). According to this theory, data instances are created by a latent process that depends on hidden variables. The latent generating process dependencies are shown in Figure. 3.6.

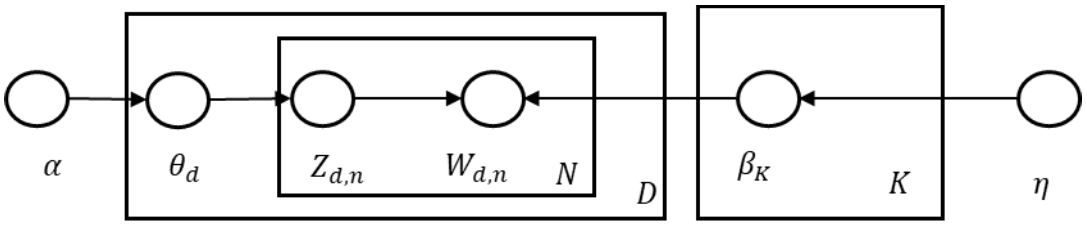


Figure 3.6. Dependencies in LDA.

Topic assignment $Z_{d,n}$ is determined by the per-document topic proportions θ_d, while θ_d is determined by the prior knowledge hyperparameter. The word $W_{d,n}$ is determined by the topic assignment β_k, which is determined by the hyperparameter β_k. Equation 3.2 expresses the joint probability distribution (over hidden variables) modelled from 1.

$$P(\beta_{1:K}, \theta_{1:D}, Z_{1:D,1:N}, W_{1:D,1:N}) =$$

$$\prod_{k=1}^{K} p(\beta_k \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(Z_{d,n} \mid \theta_d) p(W_{d,n} \mid \beta_{1:K}, Z_{d,n}) \right) \quad \textbf{(3.2)}$$

The probability distribution over the words is represented by each of theβ_k, and the topics are represented by$\beta_{1:K}$. For the d_{th} document, the topic distribution is θ_d, where $\theta_{1:D}$ denotes the probability distributions among subjects for every D documents. The topic assignment for each of N words in each of D texts is $Z_{1:D,1:N}$. $W_{1:D}, 1{:}N$ are the used words for each of the D number of documents. The number of topics K is a key parameter which determines the success of LDA. This research also looks into accurately evaluate the number of topics to optimise the LDA model.

3.6.1 Methods for finding the optimal number of topics in LDA

3.6.1.1 Perplexity

It's a statistical tool for determining how well a model handles fresh data that it hasn't seen previously. It is used in LDA to determine the ideal number of topics. In general, it

is thought that the lesser the perplexity value, the greater the accuracy. For an M-document test set, Perplexity (P) is defined as eq. 3.3, where $p(w_d)$ is the probability of document d observed words. Total number of words in document d is referred to as N_d as shown in eq 3.3:

$$P = \exp\left\{\frac{-\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\}. \quad \textbf{(3.3)}$$

3.6.1.2 Coherence

Coherence is a measure of how strongly an LDA model's induced topics are connected to one another (Röder et al. 2015). When LDA model infers a topic of terms from electronic item in a corpus of FDS text data, we classify the topic as an outlier. Such a subject is detrimental to achieving more precision. This is measured by coherence C. As indicated in eq. 3.4, it is expected that the higher the coherence value, the greater the possibility of attaining more accuracy from that model.

$$C = \sum_{ij} \text{ score}_{\text{UMass}}\left(w_i, w_j\right)$$
$$\text{score }_{\text{UMass}}\left(w_i, w_j\right) = \log \frac{D(w_i, w_j)+1}{D(w_i)} \cdot \quad \textbf{(3.4)}$$

where $D(w_i)$ represents the document frequency containing the word $w_i, D\left(w_i, w_j\right)$ represents document frequency containing both w_i and w_j, and D represents the total number of documents in the corpus.

3.7 Overall Methodology

The overall methodological flowchart for implementing sentiment analysis using DL models along with XAI for interpretability and LDA method is described in Figure 3.7. In first stage data is scrapped from product review site using a web scrapper tool (ParseHub). The main attributes captured from the website were customer reviews and labelled sentiments as star ratings. The “1” and “2” stars were classified as negative and “4” and “5” were classified as positive sentiments. The “3” star reviews were dropped from the dataset as the research is trying to find the complaints to solve the customer issue or find the positive reviews for rewarding staff and restaurants. The customer reviews were cleaned by reducing the noise and converting the words to lowercase. Also,

punctuations, questions marks, commas, colons, website url etc. were removed. Once the data was cleaned in first stage, in stage two it was used to train 3 DL models (LSTM, Bi-LTSM and Bi-GRU-LSTM-CNN) model. Based on the accuracy and low false negative score, the best model is selected among the DL models. In stage three, XAI methods (SHAP and LIME) are used on DL models to interpret the features on which outcome is predicted. Based on the features contributing the outcome, the DL model's prediction logic would be justified. Finally in stage 4, topic model is developed to identify key topics from positive and negative sentiments. The highest coherence score against number of topics is checked for developing the LDA model with right number of topics. The LDA model identifies the topics with the keywords contributing to it. Both positive and negative sentiment is categorized using the topics and its keywords.

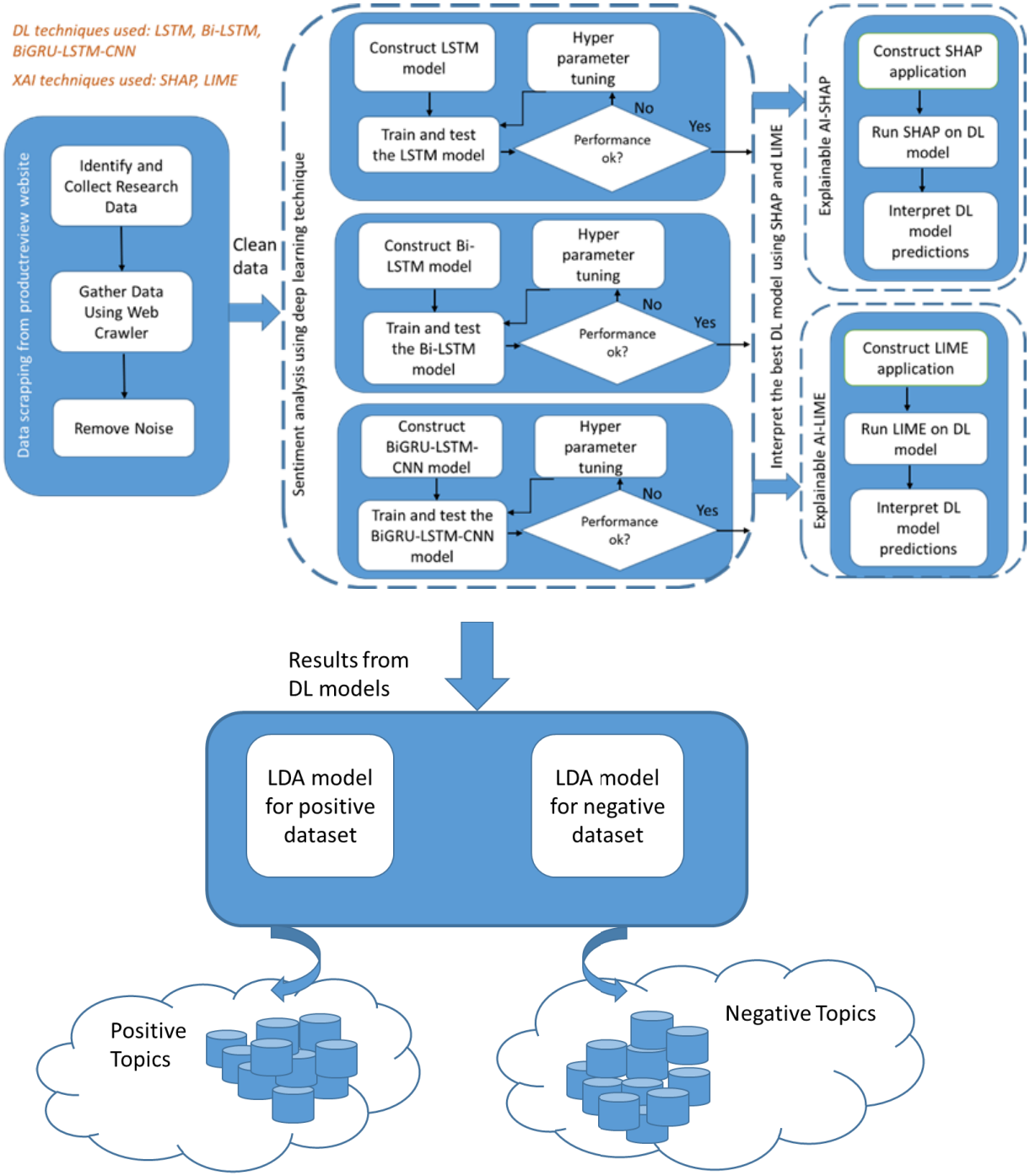


Figure 3.7. Overall Methodology flow chart with DL model, XAI technique and LDA model adopted in this work.

3.8 Implementation of the methodology

3.8.1 Objective 1

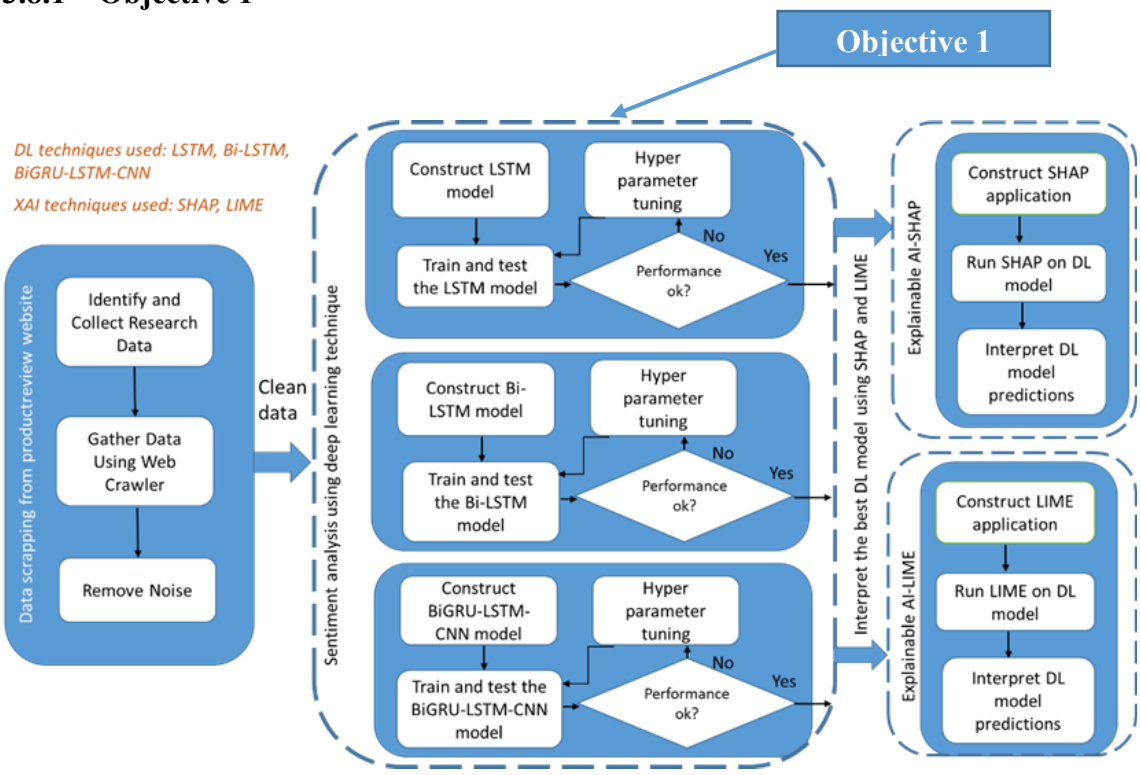


Figure 3.8. Methodology flow chart with DL technique adopted in this work.

Figure 3.8 shows the data captured from product review website. In the current study, the dataset is captured from productreview website using web scrapping tool. The dataset contains the customer reviews and star ratings from various FDS companies such as Ubereats, Menulog, Deliveroo, Youfoodz, etc. The data was cleaned to remove noise from it before using it for training DL models. Two simple DL models (LSTM and BiLSTM) and one hybrid complex model (BiGRU-LSTM-CNN) was used for performing sentiment analysis. The hyper parameters, which include epochs, batch size, layers, dropouts, number of units, and activation function, were trained and tested numerous times before being finalised. One embedding layer for word embedding, one spatialdropout1d layer for training less features, LSTM layer, flatten layer, two dense layers with the second one employing SoftMax, and one dropout layer with 50% positioned between the dense layers comprised the LSTM model. The Bi-LSTM model was created utilising one embedding layer for word embedding, one spatialdropout1d

layer for training fewer features, one Bi-LSTM layer, flatten layer, two dense layers with the second one using SoftMax, and one dropout layer with 50% positioned between the dense layers. One embedding layer for word embedding, one spatialdropout1d layer for training fewer features, one bi-directional GRU layer with two LSTMs (one forwards and one backwards), 1D convolutional layer, one global average polling 1D later and one global max pooling 1D layer, two dense layers with the last one using SoftMax, and one dropout layer with 50% located between the dense layers were used in the Bi-GRU-LSTM-CNN model. After experimenting with various combinations of hyper parameters, the models produced the best results with 100 epochs and a batch size of 32. After experimenting with various hyperparameter combinations, the models produced the best results with 100 epochs and a batch size of 32. The model was compiled with Adam optimiser (Chandriah and Naraganahalli 2021) and sparse categorical cross-entropy loss function (Mangal et al. 2019). Each of the three classifiers used 80% of the data for training and 20% for testing the models. All three DL model achieves objective 1 by performing sentiment analsyis on the FDS customer review datatset. Further using evaluation and performance metrics, the best model was picked among the three DL models.

3.8.2 Objective 2

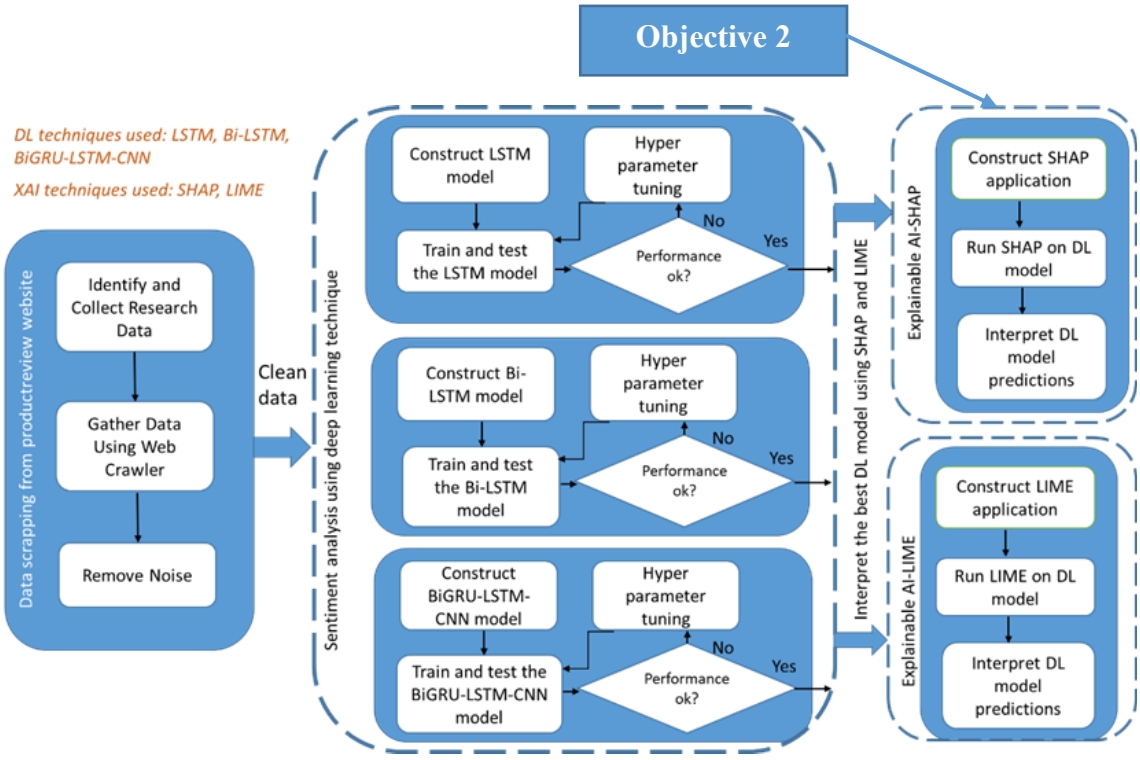


Figure 3.9. Methodology flow chart with XAI technique adopted in this work.

Figure 3.9 shows that DL models perform sentiment analysis on the raw data and then the model is validated using XAI techniques. Two different XAI techniques (SHAP and LIME) were implemented on the optimal DL model found from objective 1. SHAP methods were applied on the optimal DL model to interpret the feature importance considered by the model while predicting outcome. A DeepExplainer class from SHAP library was used to generate SHAP values for the test dataset. It took around 20 mins to generate SHAP values for the test dataset. SHAP force plot showed the feature contribution of the DL model while predicting outcome. Each arrow strip depicts how the related attribute affects the target variable's distance from or proximity to the base value. In compared to the base value, red strips indicate that their related feature pushes the value up on the higher side (showing a negative customer review), whilst blue strips indicate that the associated feature pushes the value down on the lower side (meaning a positive customer review). The LIMETextExplainer class from the LIME library was used to predict the class using variations in a probability value on the same two customer reviews utilised by SHAP before. It took only 2– 3 minutes for LIMETextExplainer to train and

generate local explanations for predictions. Using SHAP and LIME explanation method on DL model, the contributions of the words predicting the outcome can easily found and verified.

3.8.3 Objective 3

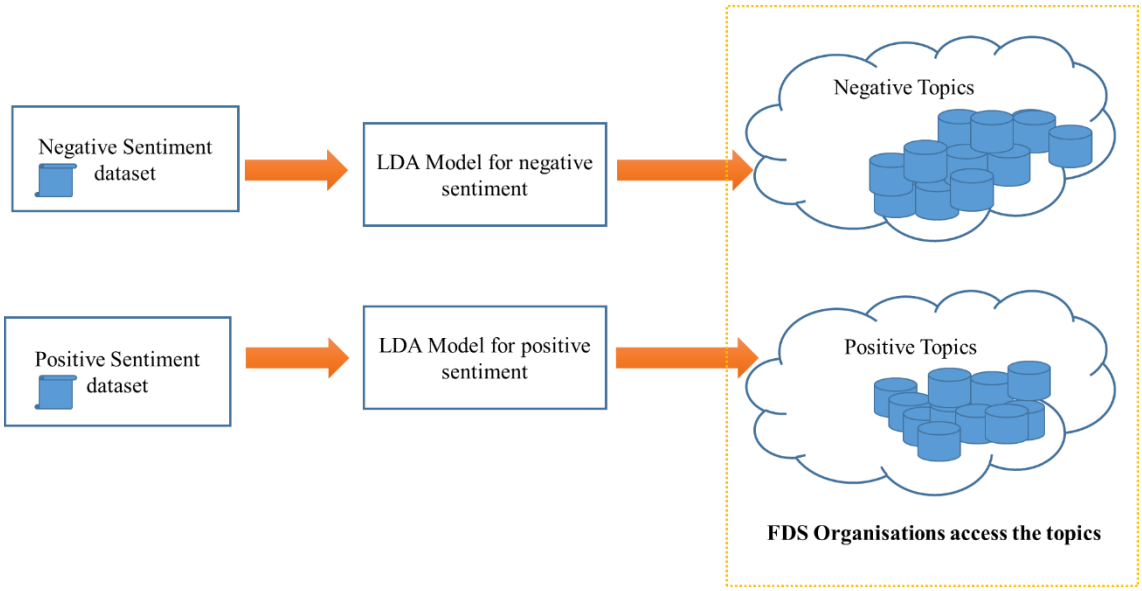


Figure 3.10. Methodology flow chart with LDA adopted in this work.

The customer review dataset was classified into positive and negative sentiments from the results of objective 1. The objective 2 validates the prediction logic of the DL model through SHAP and LIME methods. In objective 3, the negative and positive sentiment dataset is separately used to build LDA model. To build the optimal LDA model, the right number of topics is required to be predicted. To get the right number of topics, a graph is plotted on coherence score against number of topics. The highest coherence point is used to pick the optimal number of topics. On both positive and negative sets of customer review datasets, the Gensim package's LDA is used to determine the key subjects in each sentiment. Before submitting it to bigram models, the dataset is cleaned to remove noise, lemmatized (words are converted to their root words), and tokenized. The interactive chart in the pyLDAvis package is used to view the LDA model output results. The keywords connected with each category's created topics are evaluated. Topics are represented by the bubbles on the left, while important keywords are represented by the

words on the right. A strong topic model will have large, non-overlapping bubbles dispersed over the chart rather than being grouped in one quadrant. Many overlaps, such as little bubbles grouped in one area of the graph, are a sign of a model with too many topics. The topics found from positive sentiment dataset can be used for looking into the areas which is working good for FDS organisations. The topics detected from negative sentiment dataset can be used by FDS organisations to improve the problems faced by the customers.

3.8.4 Evaluation and performance metrics

To understand the accuracy of the models, the confusion matrix and F1 score of precision and recall metrics of the ML and DL models were used for comparison. The confusion matrix is one method for assessing the performance of a machine learning classification problem with two or more classes. As shown in the Table 3.2, the table has four alternative combinations of predicted and actual values.

Table 3.2. Confusion Matrix

		Actual Values: Positive	Actual Values: Negative
Predicted Values	Positive	True Positive	False Positive
Predicted Values	Negative	False Negative	True Negative

Let us understand each of the confusion matrix terms used in the research topic context. The objective is to find the negative sentiment or complaints from the review dataset.

True Positive:

The interpretation of true positive is the model has predicted the outcome as true and in actual its true case. In sentiment analysis scenario, model predicts there are complains in the dataset and actually it is true.

True Negative:

The interpretation of true negative is the model has predicted the outcome as false and in actual its true case. In sentiment analysis scenario, model predicts there are no complains in the dataset and actually its true.

False Positive (Type 1 Error):

The interpretation of false positive is the model has predicted the outcome as positive and in actual the its false case. In sentiment analysis scenario, model predicts the customer review as negative however it is actually positive.

False Negative (Type 2 Error):

The interpretation of false negative is the model has predicted the outcome as false and in actual its false case. In sentiment analysis scenario, model predicts the customer review is not negative (means positive) but actually it is negative.

In sentiment analysis, FDS organisation would like to keep the False Negative (Type 2 Error) as minimum as possible to avoid losing any customer complains.

The formulae for calculating precision eq. 3.5 and recall eq. 3.6 are as follows:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}, \quad \textbf{(3.5)}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}. \quad \textbf{(3.6)}$$

From the above equations, precision should be used when the cost of false positive for the business is more, whereas recall should be used when the cost of false negative is higher for business. The F1 score (eq. 3.7) is used to seek balance between the two metrics.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}. \quad \textbf{(3.7)}$$

The accuracy (eq. 3.8) which is the percentage of all correctly classified observations can be calculated as follows:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \quad \textbf{(3.8)}$$

Accuracy is easy to interpret as compared to $F1$ score and used when the classes are balanced and True positives and True negatives are more important. $F1$ score takes account of how the data is distributed and $F1$ score is used when the false negatives and false positives are more critical.

Also, while building LDA model, coherence score and perplexity needs to be given importance. The higher the coherence score and lower the perplexity, more accurate the LDA model becomes. In this study number of topics is decided based on the higher coherence score.

3.9 Summary

The summaries attained from the developed models were delineated in this chapter for sentiment analysis, adding explanability to black box models and topic modelling of the datasets are as follows:

1. Customer review dataset was obtained using web scrapping tool named ParseHub on Productreview website.
2. Two simple DL models (LSTM and Bi-LSTM) and one hybrid DL model (Bi-GRU-LSTM-CNN) were developed and trained using the training dataset obtained from Productreview website.
3. The DL models are compared on basis of the accuracy, low false negative score and the best model is picked for sentiment analysis.

4. For FDS organisations, trusting a highly accurate DL black box model without knowing its decision-making logic is tough.
5. XAI techniques SHAP and LIME methods are implemented on DL models to interpret the outcome on basis of word contribution.
6. Topic modelling technique, LDA model is implemented on the dataset to find relevant topic on the positive and negative dataset.

4 RESULTS AND DISCUSSION

4.1 Introduction

This chapter demonstrates the results of performing sentiment analysis on customer reviews in FDS domain using DL techniques. Three DL models are compared on the basis of the high accuracy and lower false negatives. The explanability of the black box DL models to support the prediction are also presented in this section. Finally, LDA model is implemented to select the various topic groups on which improvements can be made by FDS organisations to enhance customer satisfaction.

4.2 Results of Objective 1

4.2.1 Sentiment Analysis using simple and Hybrid DL models

This section in research tries to find which deep learning classifier would be best suited to pick FDS customer complaints from feedback and work on its solution. The research compares the accuracy and false negatives on simple DL models such as LSTM and Bi-LSTM and complex hybrid model Bi-GRU-LSTM-CNN. Based on research hypothesis 1, research expects the DL models to come up with high accuracy for sentiment prediction. Based on the success of DL models such as LSTM and Bi-LSTM, the research used these models with additional hybrid model for evaluation. It is also expected from the literature and hypothesis that the DL models would have no interpretability which can explain the outcome logic for computation.

4.2.2 Discussion

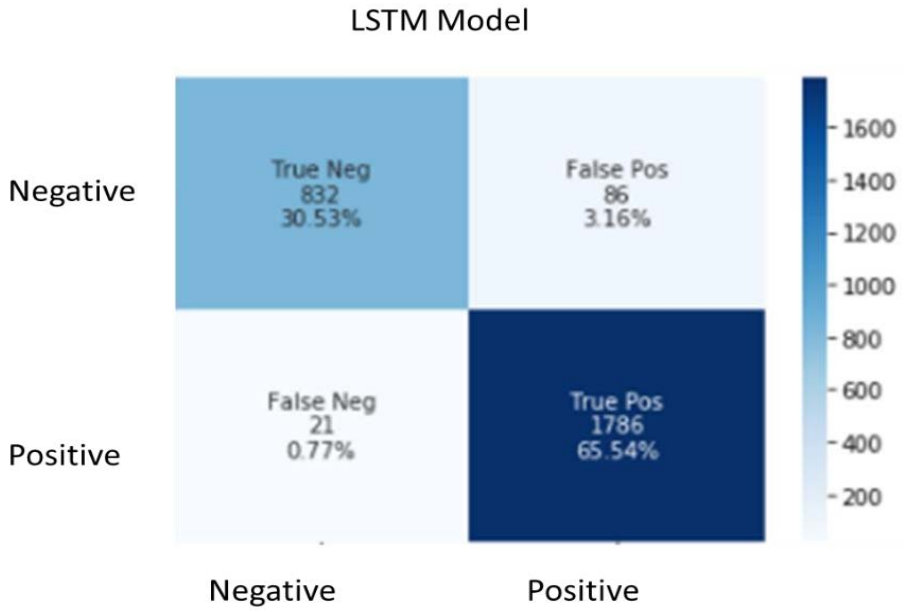
The study was able to solve the blackbox nature of DL methods (Luo and Xu 2021) by implementing XAI techniques such as SHAP and LIME. This study experimented with the ProductReview website dataset of various FDS organisations, such as Menulog, Deliveroo, Uber Eats and Youfoodz across Australia, so that it covers all the locations across Australia. This was one of the limitation identified by Luo and Xu (2021) in their research work to test the robustness of the DL model across different restaurant locations. The customer review FDS dataset was collected from ProductReview website and

cleaned to remove noise before using it for training DL models. The results of the DL models were checked by tweaking the hyper parameters after multiple rounds of training and testing. The DL models were trained and tested several times before finalising the hyperparameters, which include epochs, batch size, layers, dropouts, number of units, and activation function. The LSTM model was built with one embedding layer for word embedding, one spatialdropout1d layer for training lesser number of features, LSTM layer, flatten layer, two dense layer with the second one using SoftMax, and one dropout layer with 50% located between the dense layer. One embedding layer for word embedding, one spatialdropout1d layer for training fewer features, one Bi-LSTM layer, flatten layer, two dense layers with the second one using SoftMax, and one dropout layer with 50% located between the dense layers were used to create the Bi-LSTM model. The Bi-GRU-LSTM-CNN model was developed with one embedding layer for word embedding, one spatialdropout1d layer for training fewer features, one bi-directional GRU layer with two LSTMs (one forward and one backward), 1D convolutional layer, one global average polling 1D later and one global max pooling 1D layer, two dense layers with the last one using SoftMax, and one dropout layer with 50% located between the dense layers.

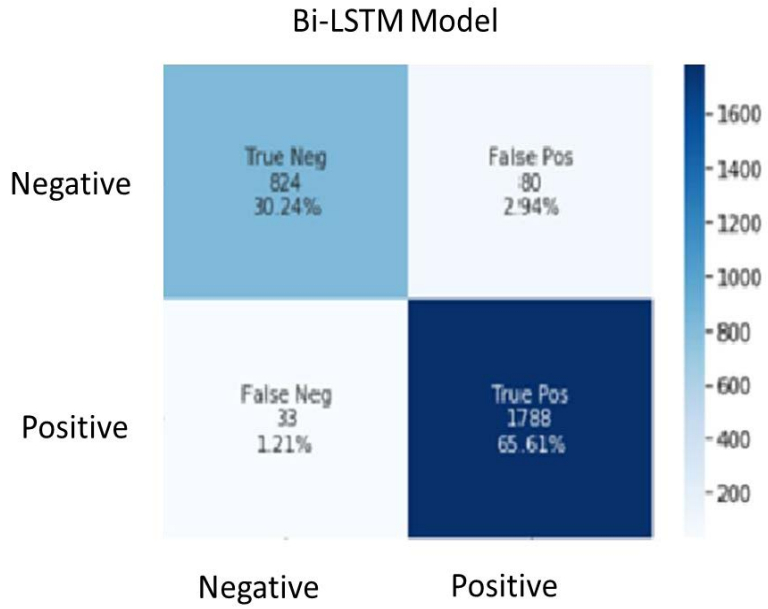
The models achieved optimum results with 100 epochs and batch size of 32 after trying with various combinations of hyperparameters. The model was compiled with Adam optimiser (Chandriah and Naraganahalli 2021) and sparse categorical cross-entropy loss function (Mangal et al. 2019). All the three classifiers considered 80% data for training and 20% for testing the models.

The classification performance and errors of the classifiers are represented in the confusion matrix. The type 1 error is shown by false positives, and type 2 is shown by false negatives. The significance of an error is determined by the classification problem's domain. In the case of FDS, higher importance will be given to type 2 errors. Type 1 error denotes that the alert raised for positive customer review comments as complaint, which will require some operational effort to investigate and close the customer comment as not a complaint. Type 2 error indicates that the system cannot identify the negative sentiments, which is a larger risk because the customer complaints will not be detected by the system. The FDS organisations prefer to identify and work on each and every

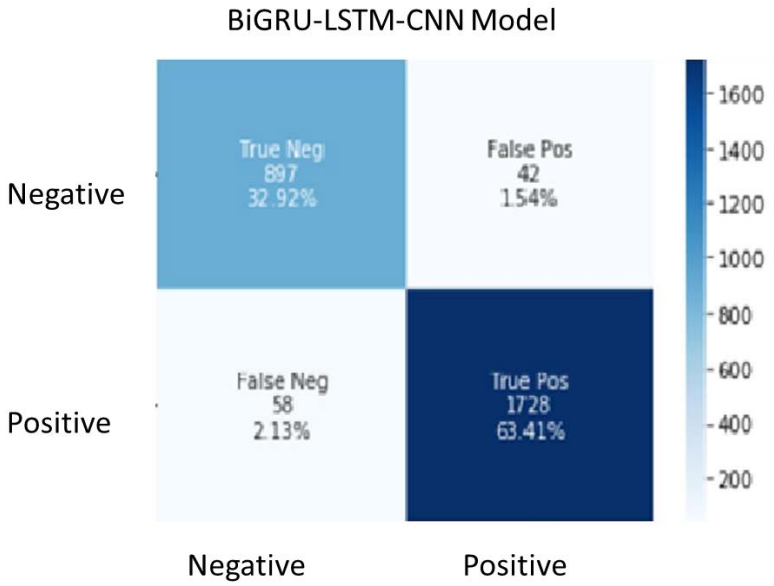
customer complaint to improve customer satisfaction. Hence, the model should have lesser false negatives in its prediction. The false negatives can be computed from confusion matrix of the DL models. Figure 4.1 shows the confusion matrix generated from LSTM, Bi-LSTM and Bi-GRU-LSTM-CNN model.



(a)



(b)



(c)

Figure 4.1. Confusion matrix of (a) LSTM; (b) Bi-LSTM; and (c) Bi-GRU-LSTM-CNN model

The confusion matrix shown in Figure 4.1. (a) clearly indicates that the LSTM classifier can perform accurate prediction (65.54% reviews, which are positive, and 30.53% reviews, which are negative), achieving an overall accuracy of 96.07%. Only 0.77% reviews give false negative results, whereas 3.16% returns false positive results. The numbers from the confusion matrix are validated with the performance metrics (Table 2[a]) by using the assessment measures.

Similarly, the confusion matrix (Figure 4.1. [b]) shows that the Bi-LSTM classifier can perform accurate prediction (65.61% reviews, which are positive, and 30.24% reviews, which are negative), resulting in a 95.85 percent overall accuracy. The Bi-LSTM classifier gives 1.21% false negative results, and 2.94% returns false-positive results. The numbers from the confusion matrix are validated with the performance metrics (Table 4.1[b]) by using the assessment measures.

The confusion matrix (Figure 4.1. [c]) shows that the Bi-GRU-LSTM-CNN classifier can perform accurate prediction (63.41% reviews, which are positive, and 32.92%, reviews which are negative), resulting in a 96.33% overall accuracy. The Bi-GRU-LSTM-CNN

classifier gives 2.13% false negative results, and 1.54% returns false-positive results. The numbers from the confusion matrix are validated with the performance metrics (Table 4.1.[c]) by using the assessment measures.

Table 4.1. Performance metrics - (a) LSTM; (b) Bi-LSTM; and (c) Bi-GRU-LSTM-CNN model.

(a)

	Precision	**Recall**	**F1_score**	**OA**
Negative	0.98	0.91	0.94	96.07
Positive	0.95	0.99	0.97	

	Precision	**Recall**	**F1_score**	**OA**
Negative	0.96	0.91	0.94	95.85
Positive	0.96	0.98	0.97	

(b)

	Precision	**Recall**	**F1_score**	**OA**
Negative	0.94	0.96	0.95	96.33
Positive	0.98	0.97	0.97	

(c)

The results from the above performance metrics show that all the DL models developed for performing sentiment analysis attain high overall accuracy (LSTM at 96.07%, Bi-LSTM at 95.85%, and Bi-GRU-LSTM-CNN at 96.33%). However, FDS organisations will pick the LSTM model as the best classifier due to its lesser type 1 error with 21 false negatives as compared to BiLSTM with 33 and Bi-GRU-LSTM-CNN with 58.

Table 4.1 in results section shows that LSTM, Bi-LSTM and Bi-GRU-LSTM-CNN obtain an accuracy of 96.07%, 95.85% and 96.33%, respectively. The DL models

achieved higher accuracy as compared to the models developed in the past in other research works. Table 4.2 shows the accuracy achieved in DL/Ml models predicting customer sentiments in FDS domain from recent papers. However all the ML/DL methods used in the past are not interpretable.

Table 4.2. Accuracy scores achieved in ML/DL models from recent papers.

Method	Accuracy	Interpretable	DL/ML	References
Random Forest	89%	No	ML	(Luo and Xu 2021)
GBDT	87.5%	No	ML	(Luo and Xu 2021)
Simple Embedding + Average Pooling	91.1%	No	DL	(Luo and Xu 2021)
Bidirectional LSTM	90.8%	No	DL	(Luo and Xu 2021)
SVM	91.5%	No	ML	(Adak et al. 2022)

Comparing the previous work done in research, we found that the DL models implemented in this research have acquired higher accuracy. The next task is to find out the best DL model in terms of accuracy from the research work.

Although the accuracy of the LSTM model is high, it lacks model interpretability and explainability of the decisions made. The explanations of the LSTM-based black box model will help build the trust in the system.

4.2.3 Validation

The research shows that the LSTM model is effectively used to identify the positive and negative sentiments from customer reviews with an accuracy of 98.07% with lesser false negative rate of 0.77%. The LSTM model is validated with precision, recall and f1_score. Further, it will be examined for interpretability using XAI techniques in the next section.

4.3 Results of Objective 2

4.3.1 XAI explanation on LSTM model using SHAP and LIME

The research shows from previous steps, DL models is able to provide 96.07% accuracy through LSTM model in performing sentiment analysis. However, lack of explainability of the DL models is an issue for the industry to trust for its use in market. The research hypothesis 2 expects XAI technique to uncover the prediction logic of LSTM model successfully. XAI methods should be able to show the word contributions in the customer review to show the sentiment prediction. The research hypothesis would falsify if the prediction logic of the LSTM model is not verified using XAI techniques. Therefore, XAI techniques SHAP and LIME are implemented on the DL model to interpret the features on which the outcome is predicted. Also, the output of SHAP and LIME are compared in this section.

4.3.2 Discussion

SHAP was implemented on the model to interpret the feature importance considered by LSTM while making the predictions after training and testing the LSTM classifier. A DeepExplainer class from the SHAP library, which took approximately 20 min, was used to generate the SHAP values for the test dataset. Figures 4.2 and 4.3 show the force plot representing the interpretation of two customer review predictions made by the LSTM classifier. The base value shown on the plot is the average value of the target variable across the dataset we passed to the DeepExplainer class. Each arrow strip shows the effect of its associated feature on pushing the target variable away or close to base value. Red strips show that their associated feature pushes the value on the higher side (indicating customer review being negative) in comparison to the base value, whereas the blue strips indicate that the associated feature pushes the value down on the lower side (indicating customer review being positive).

Figure 4.2 represents the SHAP explanation for the LSTM model's detection of a positive customer review. The inference from the force plot and customer review suggests that the customer is very happy with the customer service for getting the new delivery of the meal after requesting it because the customer was on crutches. The words represented in blue colour contribute to positive sentiment, and the words shown in red colour contribute to

negative sentiment. The explainable model showed that words, such as ‘impressed’, ’new’ and ‘have’, strongly pushed the output prediction value to positive sentiment, which matches with the actual positive customer review prediction.

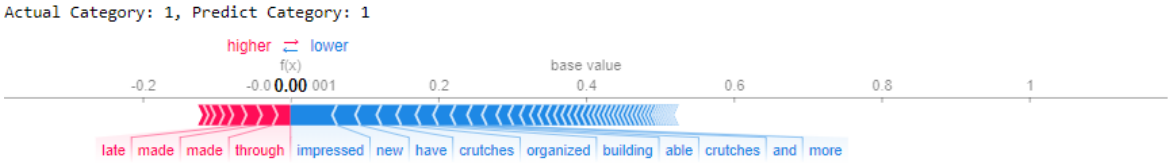


Figure 4.1. SHAP explanation on the positive customer review.

Figure 4.3 shows the SHAP explanation for negative customer review prediction. The inference from the force plot and customer review is that the customer is asking for a refund because the ordered subway came without salad and sauce. The words, such as ‘refund’, ‘not’, ’why’ and ‘entitled’, show a positive correlation with the negative customer review prediction.

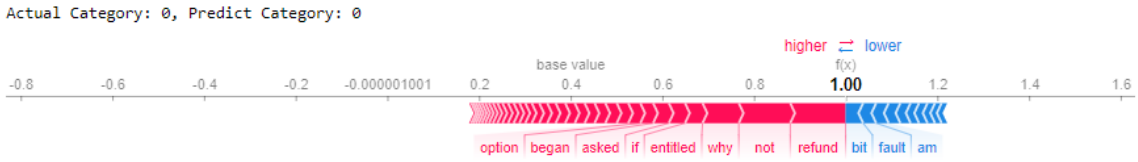


Figure 4.2. SHAP explanation on the negative customer review.

From Figures 4.2 and 4.3, the LSTM model using the SHAP technique can be validated whether the right words contribute to the right prediction. The SHAP interpretation identifies satisfactory reasoning for the predictions made by the LSTM model. It gives a good insight into the FDS organisations to decide if the identified negative customer review is a false positive and requires further investigation by inspecting these indicators along with the actual meaning of the customer reviews.

LIMETextExplainer class from the LIME library was used to predict the class with the variations of a probability value on the same two customer reviews previously used by SHAP. LIMETextExplainer took only 2–3 min to train and generate local explanations for predictions.

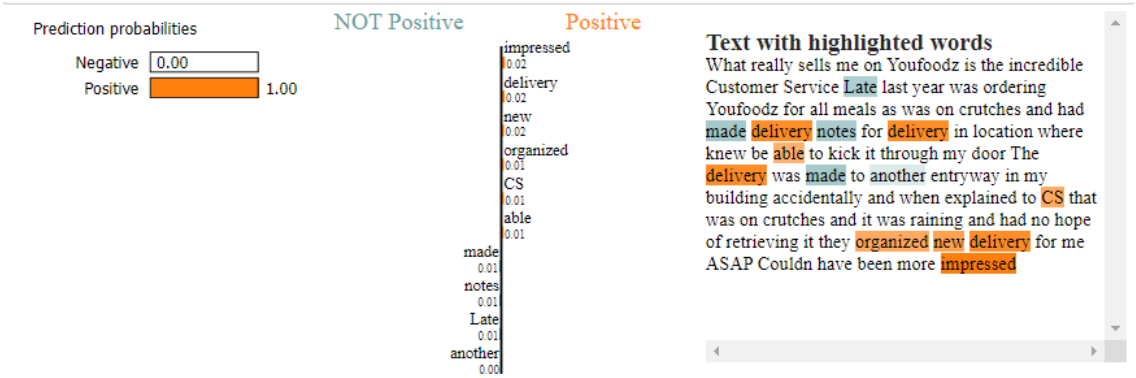


Figure 4.3. LIME explanation on the positive customer review detected by the LSTM model.

For customer review 1 in Figure 4.4, the LSTM model is 100% certain that the review is positive sentiment. The words, such as 'impressed', 'delivery' and 'new', increased the review's chance to be classified as positive. However, the feature contribution of the positive words classifying the customer review as positive looks similar in the LIME explainer graph.

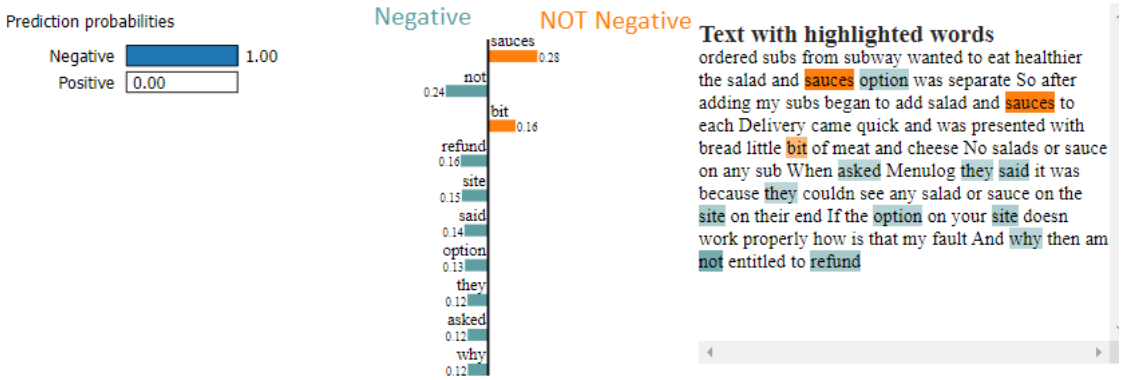


Figure 4.4. LIME explanation on the negative customer review detected by the LSTM model.

In the next example on customer review 2, the LSTM model is 100% certain that the customer review (shown in Figure 4.5) is negative sentiment. LIME explainer suggests that the words, such as 'not', 'refund' and 'site', show a positive correlation with the negative customer review prediction.

4.3.3 Validation

SHAP and LIME allowed us to perform an in-depth analysis of the model with its sample customer review test data. For positive customer review, SHAP and LIME picked key feature words, such as ‘impressed’, ’new’, ’delivery’ and ’have’, which strongly pushed the output prediction value to positive sentiment, which matches with the actual positive customer review prediction. The feature contribution of the positive words classifying the customer review looks flat in the LIME explainer graph as compared to the SHAP force plot graph. Similarly, for negative customer sentiment review, the explainers suggested that words, such as ’not’, ’refund’, ’why’ and ’site’, show a positive correlation with the negative customer review prediction. SHAP’s ability to show the interpretation of LSTM predictions by pinpointing the contribution score of each feature is better as compared to that of LIME. However, SHAP took more time to train with the dataset compared with LIME.

4.4 Results of Objective 3

4.4.1 Topic Categorization of negative and positive sentiments using LDA

The FDS organisations with the use of DL models along with XAI methods is able to classify negative and positive sentiments from the previous objectives. The next step for the business is to know the issues or problems which can be further improved. Similarly, they would like to know the things which are being appreciated by the customers. To solve the problem of topic modelling from the customer reviews, Latent Dirichlet Allocation (LDA) technique is implemented. The LDA model picks the topics on which the maximum customers are complaining or appreciating which can be further taken by the FDS organisations to solve issues or reward staffs. According to research hypothesis 3, the research expects the Topic modelling technique to categorize the negative customer reviews into various topics/categories which can be later sent to proper department within FDS organisation to solve the issue. Similarly according to research hypothesis 4, the research expects the Topic Modelling method to classify the positive topics which can be further used to reward staffs and restaurants. In case, if LDA model fails to classify topics from the given dataset, the research hypothesis 3 and 4 would fail.

4.4.2 Discussion

From objective 1 and objective, two sets of sentiments (positive and negative) customer reviews are classified. The Latent Dirichlet Allocation (LDA) from Gensim package is implemented on both positive and negative sets of customer review dataset to extract the key topics in each sentiment. The dataset is cleaned to reduce noise, lemmatized (converting word to its root word) and tokenized before passing it to bigram models. The LDA model is created by passing the dictionary (id2word) and the corpus. The other hyperparameters passed are “alpha=auto”, “eta” which affect the sparsity of the topics, “chunksize = 100”, update_every=1”,”passes=10” and num_topics = “10”. The below Table 4.3 shows LDA model with negative and positive customer reviews is built with 10 different topics where each topic is combination of keywords and each keyword contributes a certain weightage to the topic.

Table 4.3. Word contribution for topic on (a) negative and (b) positive reviews.

Topic Number	Combination of Keywords with weights
Topic 1	0.272*"use" + 0.061*"restaurants" + 0.053*"long" + 0.019*"hrs" + ' '0.019*"zero" + 0.016*"information" + 0.014*"occasions" + 0.014*"costs" + ' '0.012*"total" + 0.011*"noticed"
Topic 2	'0.103*"took" + 0.048*"correct" + 0.036*"gave" + 0.034*"try" + 0.032*"via" + ' '0.025*"name" + 0.022*"site" + 0.021*"manager" + 0.020*"longer" + ' '0.020*"large"
Topic 3	'0.051*"money" + 0.034*"uber" + 0.031*"credit" + 0.030*"account" + ' '0.027*"rang" + 0.026*"eats" + 0.023*"app" + 0.022*"picked" + 0.020*"menu" + '

	'0.020*"email"
Topic 4	'0.046*"order" + 0.035*"food" + 0.023*"deliveroo" + 0.022*"service" + ' '0.020*"delivery" + 0.018*"time" + 0.016*"driver" + 0.016*"restaurant" + ' '0.015*"never" + 0.014*"ordered"
Topic 5	0.080*"worst" + 0.063*"experience" + 0.045*"far" + 0.037*"guys" + ' '0.025*"eating" + 0.023*"big" + 0.022*"busy" + 0.022*"showing" + ' '0.019*"close" + 0.016*"declined"
Topic 6	0.032*"meal" + 0.024*"cancelled" + 0.021*"chat" + 0.019*"contact" + ' '0.018*"meals" + 0.017*"saying" + 0.015*"also" + 0.014*"hungry" + ' '0.014*"live" + 0.014*"cancel"
Topic 7	0.060*"sorry" + 0.043*"already" + 0.039*"past" + 0.039*"decided" + ' '0.039*"mistake" + 0.035*"local" + 0.027*"deleted" + 0.026*"months" + ' '0.025*"happening" + 0.025*"fast"
Topic 8	0.099*"bad" + 0.047*"let" + 0.036*"care" + 0.031*"road" + 0.026*"less" + ' '0.022*"riders" + 0.020*"soggy" + 0.017*"pretty" + 0.016*"provided" + ' '0.016*"etc"
Topic 9	0.081*"instead" + 0.066*"paid" + 0.060*"happy" + 0.041*"chicken" + ' '0.018*"chef" + 0.018*"old" + 0.017*"pizzas" + 0.017*"muscle" + '

	'0.014*"frustrating" + 0.014*"complain"
Topic 10	0.045*"least" + 0.030*"canceled" + 0.029*"shop" + 0.028*"stars" + ' '0.024*"dropped" + 0.023*"burgers" + 0.021*"inedible" + 0.018*"ruined" + ' '0.018*"simple" + 0.017*"felt"

(a)

Topic Number	**Combination of Keywords with weights**
Topic 1	0.111*"muscle" + 0.061*"taste" + 0.047*"like" + 0.019*"flavour" + ' '0.017*"months" + 0.017*"plan" + 0.016*"full" + 0.016*"value" + ' '0.016*"frozen" + 0.015*"without"
Topic 2	0.066*"recommended" + 0.048*"follow" + 0.039*"friends" + 0.037*"year" + ' '0.034*"sure" + 0.031*"due" + 0.029*"problem" + 0.027*"overall" + ' '0.023*"comes" + 0.020*"works"
Topic 3	0.060*"meals" + 0.033*"food" + 0.026*"great" + 0.022*"meal" + 0.020*"chef" ' '+ 0.019*"time" + 0.018*"good" + 0.015*"delivery" + 0.013*"order" + ' '0.012*"easy"
Topic 4	0.082*"im" + 0.052*"mmc" + 0.044*"protein" + 0.034*"guys" + 0.033*"chicken" ' '+ 0.031*"still" + 0.030*"looking" + 0.027*"able" + 0.024*"vegan" + ' '0.017*"dishes"

Topic 5	0.084*"could" + 0.041*"satisfied" + 0.033*"point" + 0.032*"carb" + ' '0.024*"feeling" + 0.021*"please" + 0.019*"review" + 0.019*"kitchen" + ' '0.018*"discount" + 0.017*"store"
Topic 6	0.055*"first" + 0.042*"ordered" + 0.038*"team" + 0.037*"find" + ' '0.034*"helpful" + 0.032*"last" + 0.028*"bit" + 0.024*"friendly" + ' '0.023*"didnt" + 0.021*"box"
Topic 7	0.052*"macros" + 0.036*"give" + 0.034*"staff" + 0.028*"partner" + ' '0.023*"loss" + 0.023*"driver" + 0.021*"hard" + 0.021*"foods" + 0.020*"bad" ' '+ 0.020*"another"
Topic 8	0.129*"companies" + 0.077*"program" + 0.053*"small" + 0.033*"etc" + ' '0.031*"poor" + 0.030*"stick" + 0.024*"veg" + 0.023*"email" + 0.023*"unlike" ' '+ 0.021*"following"
Topic 9	0.156*"service" + 0.093*"customer" + 0.033*"prep" + 0.029*"never" + ' '0.028*"fantastic" + 0.026*"excellent" + 0.022*"got" + 0.021*"actually" + ' '0.019*"awesome" + 0.015*"away"
Topic 10	0.086*"definitely" + 0.052*"sizes" + 0.036*"people" + 0.033*"chefgood" + ' '0.029*"product" + 0.028*"theyre" + 0.023*"add" + 0.023*"issues" + '

	'0.022*"challenge" + 0.021*"reasonable"

(b)

Each row represents a topic with weightage of each keyword using lda_model.print_topics(). The topic 1 to topic 10 can be renamed to logical category based on the combination of weights and keywords. For example, topic 3 contains words such as “money”, ”uber”, ”credit”, ”account”, ”rang”, ”email”, ”app” etc. Topic 3 can be categorized to accounts, as it talks more towards account related issue. On the other hand, topic 4 contains "order" , “food”, “service”, “delivery”, “ time”, “driver”, “restaurant”, “ordered” and “never”. Topic 4 can be more categorized towards the delivery related issue. Topic 5 contains words like “worst”, “experience”, “far”, “guys”, “busy”, “close” and “declined” which more points towards bad customer service. The below Table 4.4 represents the identified topic names based on the keywords and its weights.

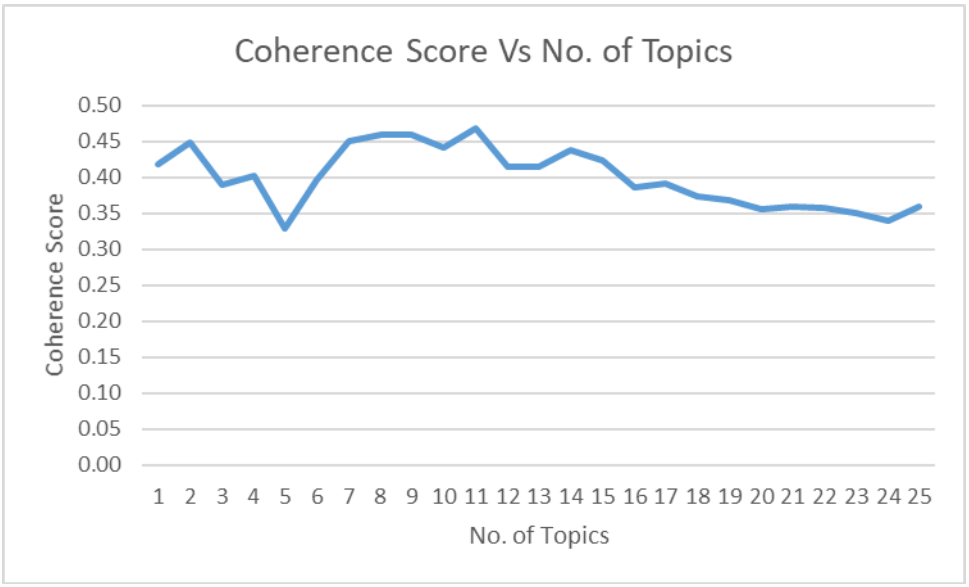
Table 4.4. Category Names derived from Keywords with weights

Category Name	Negative keywords with weights
Account	'0.051*"money" + 0.034*"uber" + 0.031*"credit" + 0.030*"account" + ' '0.027*"rang" + 0.026*"eats" + 0.023*"app" + 0.022*"picked" + 0.020*"menu" + ' '0.020*"email"
Delivery	'0.046*"order" + 0.035*"food" + 0.023*"deliveroo" + 0.022*"service" + ' '0.020*"delivery" + 0.018*"time" + 0.016*"driver" + 0.016*"restaurant" + ' '0.015*"never" + 0.014*"ordered"
Order	0.080*"worst" + 0.063*"experience" + 0.045*"far" + 0.037*"guys" + ' '0.025*"eating" + 0.023*"big" + 0.022*"busy" + 0.022*"showing" + ' '0.019*"close" + 0.016*"declined"

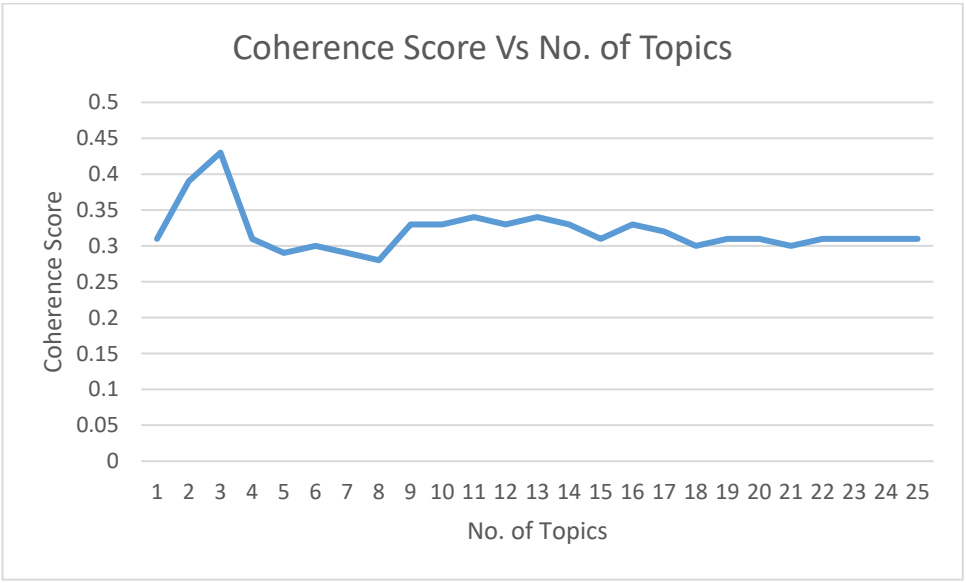
Food Quality	0.045*"least" + 0.030*"canceled" + 0.029*"shop" + 0.028*"stars" + ' '0.024*"dropped" + 0.023*"burgers" + 0.021*"inedible" + 0.018*"ruined" + ' '0.018*"simple" + 0.017*"felt"
Online customer service	0.032*"meal" + 0.024*"cancelled" + 0.021*"chat" + 0.019*"contact" + ' '0.018*"meals" + 0.017*"saying" + 0.015*"also" + 0.014*"hungry" + ' '0.014*"live" + 0.014*"cancel"

The FDS organisation can look into the negative categories along with keywords to solve the customer issues. Different teams for each category identified can be assigned to look into actual issues causing problem. Also, the categories can be prioritised by FDS organisations for resolving issues. The resolution team can look into the weights of the words to understand the detailed root cause.

Not all the topics can be easily mapped to categories and hence it is essential to know as how many topics can be extracted from given dataset. According to Hasan et al. (2021), the efficacy if LDA depends on its key parameter “number of topics” which is dependent on the dataset. Since it is dependent on dataset, the “number of topics” parameter will be different in every case. Model perplexity and topic coherence solve the problem of judging the model. Higher coherence score and lower perplexity are trusted for predicting the optimal number of topics in LDA. The research tried to find the optimum LDA model by looking in to coherence score of number of topics from 1 to 25. The below Figure 4.6 plots the graph between number of topics and coherence score.



(a)



(b)

Figure 4.5. Coherence score vs no. of topics on (a) negative (b) positive reviews.

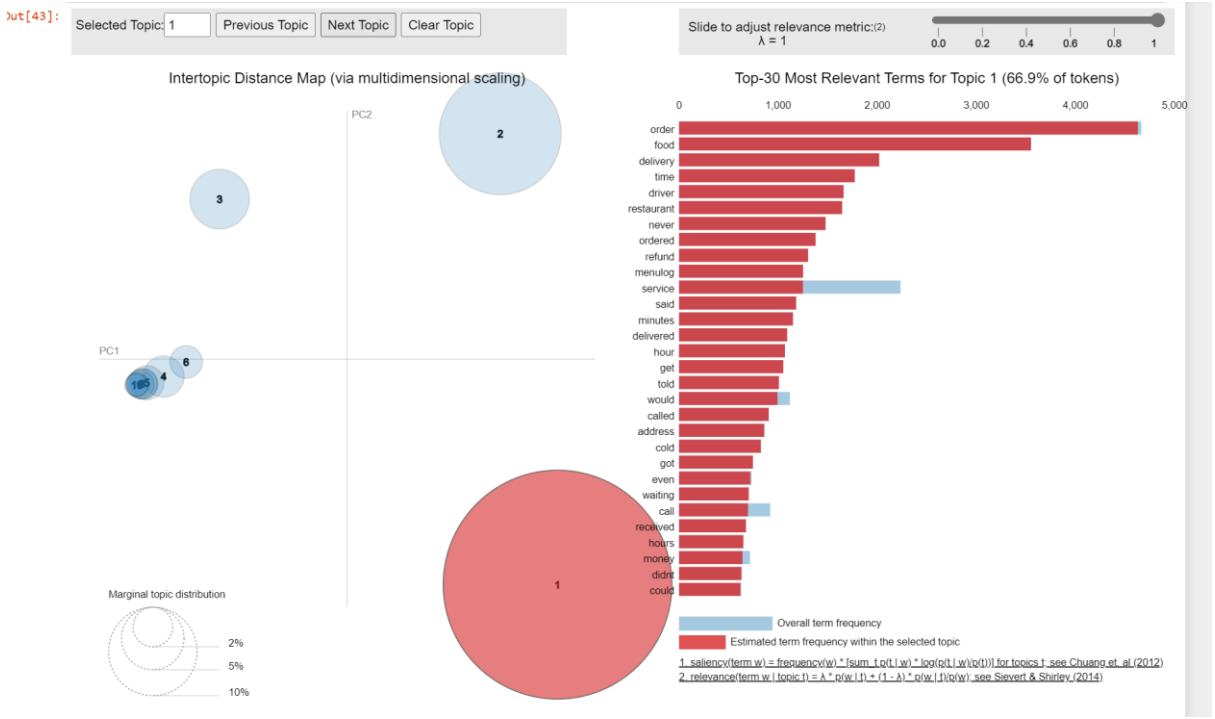
From Figure 4.6 (a), the optimum number of topics is found to be 11 as the coherence score looks to peak there before flattening out. Similarly for positive reviews, the optimum number of topics is found out to be 3, as coherence score peaked at 3 in Figure 4.6 (b) before flattening out. The research considered the below mentioned number of

topics optimised for the LDA model as shown in Table 4.5 and found the respective coherence score and perplexity.

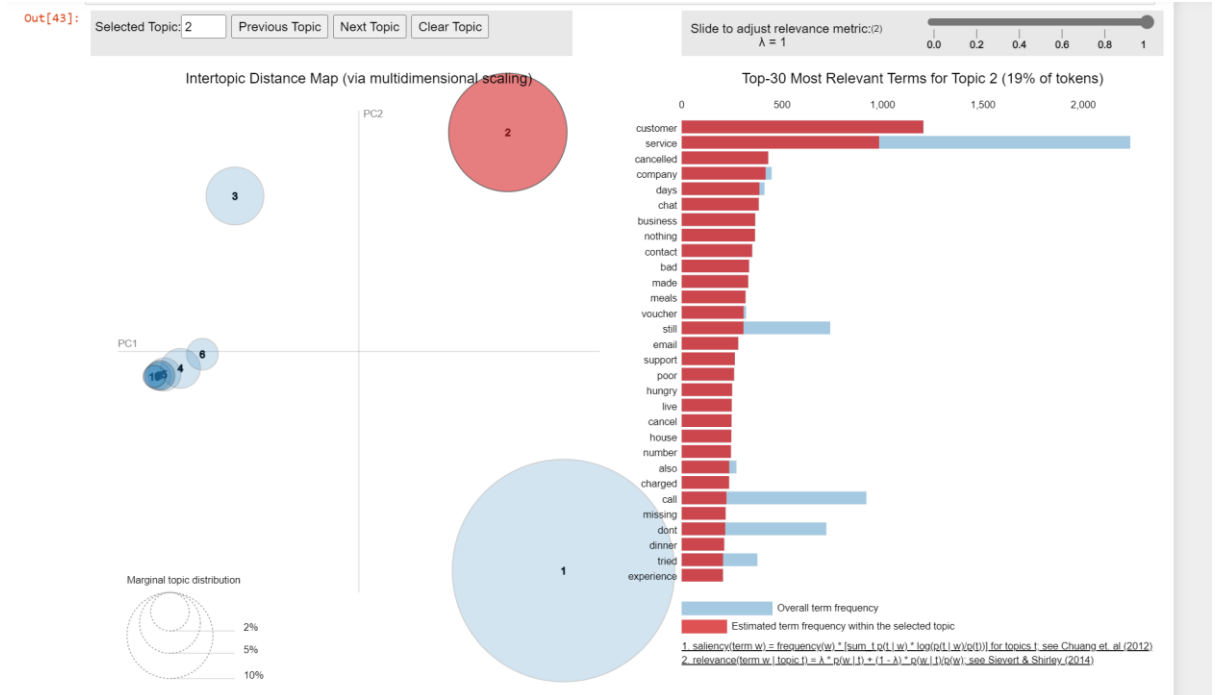
Table 4.5. Coherence score and perplexity for no. of topics.

Sentiment	No. of Topics	Coherence Score	Perplexity
Negative	11	0.4688	-8.07
Positive	3	0.4269	-7.11

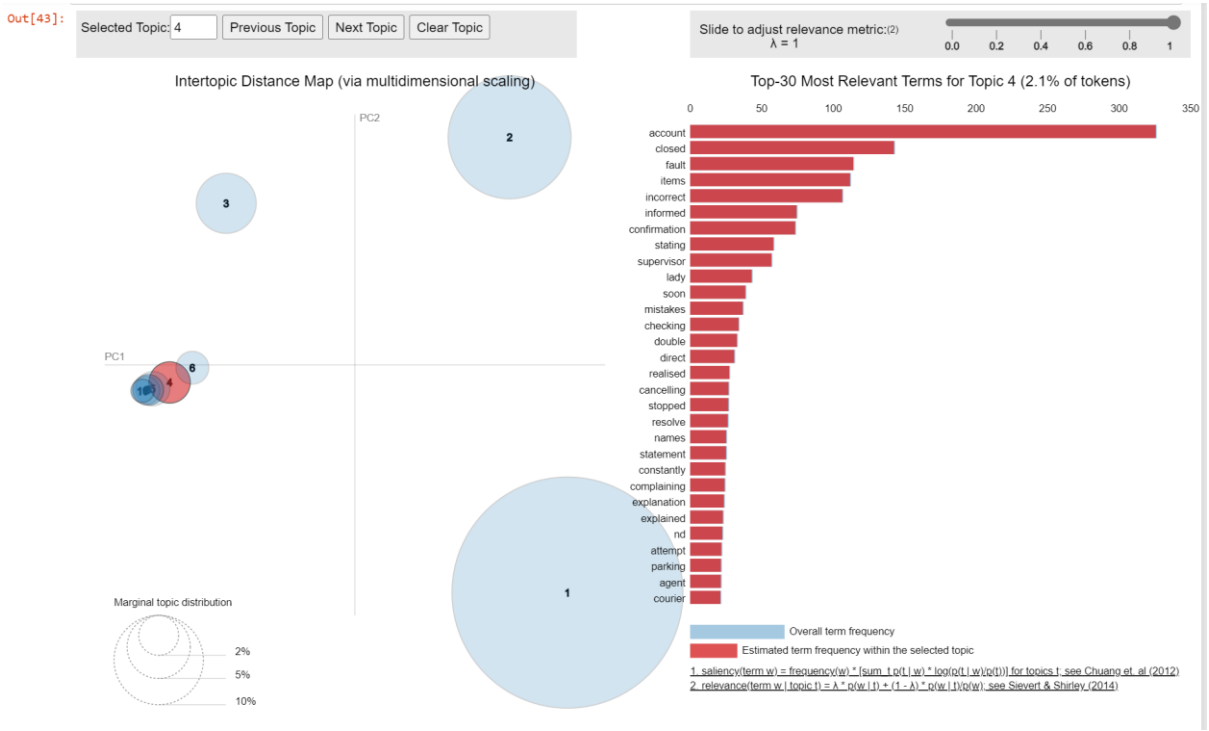
To examine the produced results by the LDA models, pyLDAvis packages's interactive chart is used. The produced topics for each category are examined with the associated keywords. The bubbles on the left side represents topics and words on the right side are salient keywords from the selected topic. Instead of being clustered in one quadrant, a good topic model will have fairly large, non-overlapping bubbles scattered throughout the chart. Many overlaps, small sized bubbles clustered in one region of the chart, are typical indication of a model with too many topics. The Figure 4.7 (a) suggests topic 1 showing keywords "order", "food", "delivery", "time", "restaurant", "never", "ordered", "refund" etc. Looking at the keywords, topic1 suggests that topic can be grouped into "Delivery" as it focuses on keywords "order", "delivery"," time", "refund". The Figure 4.7 (b) shows topic 2 having keywords "customer", "service", "cancelled", "company", "days", "chat", "business", "nothing", "contact" etc. which suggests the topic 2 is related to "Customer Service". Topic 4 in Figure 4.7 (c) shows many overlaps, small sized bubbles clustered in one region of the chart, which suggests a model with too many topics. The keywords such as "account", "closed", "fault", "items", "incorrect", "confirmation" etc. doesn't focus on one topic and it is difficult to group to relevant topic.


Out[43]:
Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (66.9% of tokens)
0 1,000 2,000 3,000 4,000 5,000
order
food
delivery
time
driver
restaurant
never
ordered
refund
menulog
service
said
minutes
delivered
hour
get
told
would
called
address
cold
got
even
waiting
call
received
hours
money
didnt
could
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

(a)


Out[43]:
Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0 0.2 0.4 0.6 0.8 1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (19% of tokens)
0 500 1,000 1,500 2,000
customer
service
cancelled
company
days
chat
business
nothing
contact
bad
made
meals
voucher
still
email
support
poor
hungry
live
cancel
house
number
also
charged
call
missing
dont
dinner
tried
experience
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

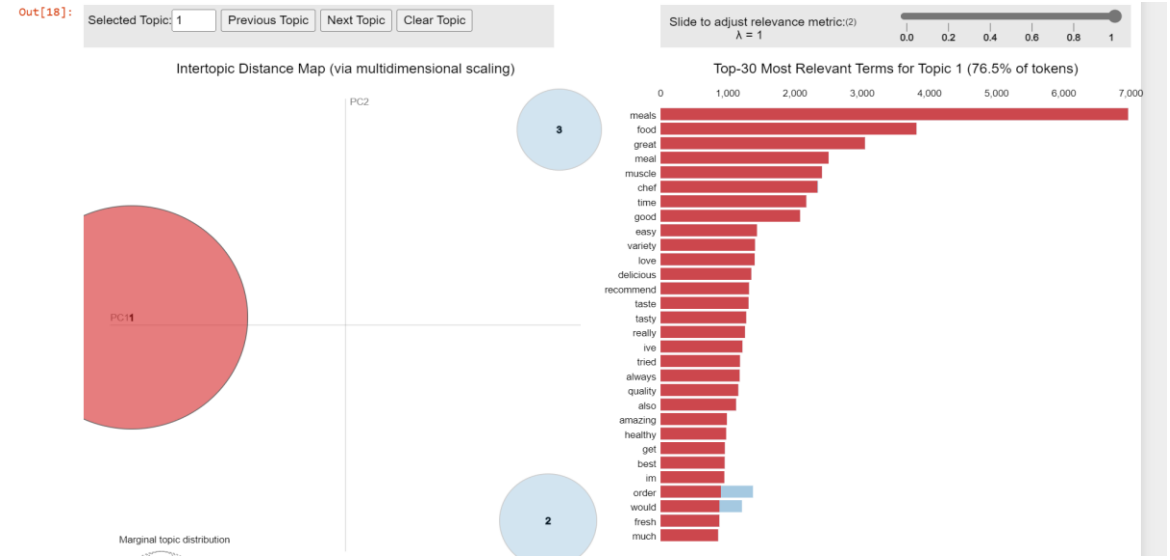
(b)



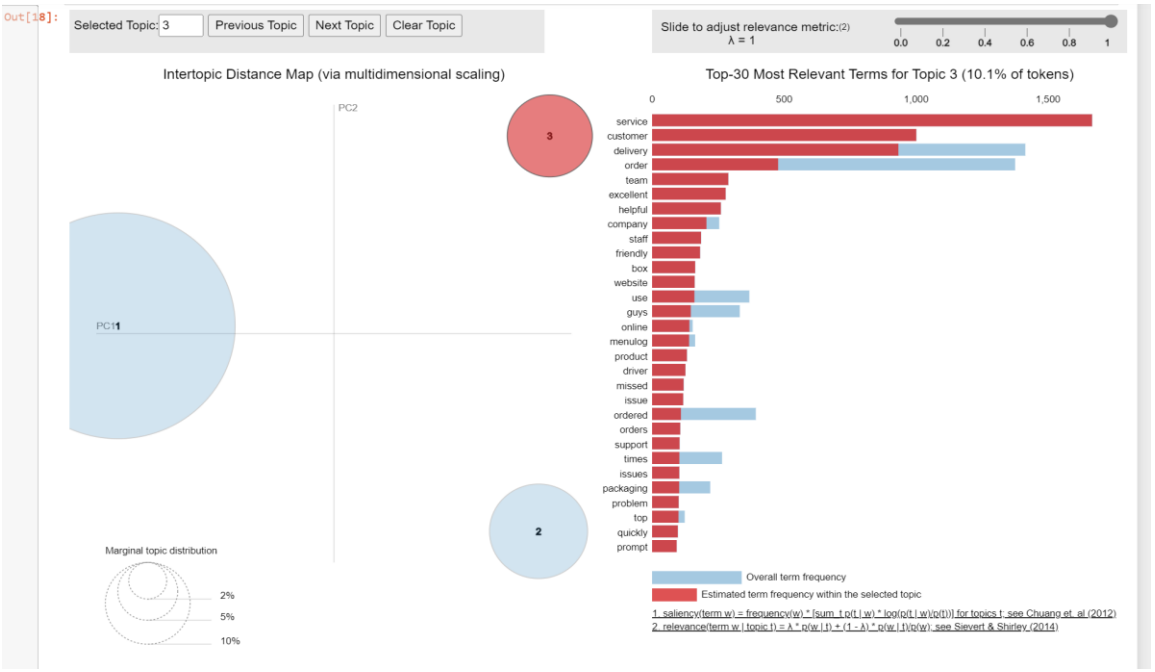
(c)

Figure 4.6. (a) (b) (c) Topics with keywords for negative sentiments.

Similarly for positive reviews, Figure 4.8 (a) shows keywords “meals”, “food”, “great”, “meal” etc. suggests topic 1 can be grouped into “Food Quality”.



(a)



(b)

Figure 4.7. (a) (b) Topics with keywords for positive sentiments.

The Figure 4.8 (b) shows the keywords “service”, ”customer”, ”delivery”, “order”, “team” ,”excellent” suggests topic 3 talks about “good customer service”.

4.4.3 Validation

The research questions pertaining to objective 3 raised question whether topic modelling technique like LDA can find various topics from the customer reviews. The research shows that with the help of high coherence score, the optimum number of topics can be predicted. Using the optimum number of topics passed into LDA model, the model predicted the topics based on keywords. The FDS customer complaints can be grouped into different positive and negative topics as shown in the Table .

Table 4.6. Positive and Negative Categories extracted from customer reviews

Negative Topics Found	**Positive Topics Found**
Account	Food Quality
Delivery	Customer Service
Order	
Food Quality	
Online customer service	

The positive reviews revealed they can be grouped into “Food Quality” and “Customer Service”. The negative reviews can be categorized into “Account”, ”Delivery”, ”Order”, ”Food Quality”, “Online customer service”. The detailed level information on the topics can be found by looking into the keywords for each topic.

4.5 Summary

In chapter 4, the application of the DL models along with XAI methods and topic modelling lead to the following results:

1. The DL models developed for performing sentiment analysis attain high overall accuracy (LSTM at 96.07%, Bi-LSTM at 95.85%, and Bi-GRU-LSTM-CNN at 96.33%).
2. FDS organisations will pick the LSTM model as the best classifier due to its lesser type 1 error with 21 false negatives as compared to BiLSTM with 33 and Bi-GRU-LSTM-CNN with 58.

3. The LSTM model has a high level of accuracy, but it lacks model interpretability and explanation of the decision it makes.
4. With the use of SHAP and LIME, the research was able to conduct an in-depth study of the model using sample customer review test data.
5. When compared to LIME, SHAP's ability to display the interpretation of LSTM predictions by identifying the contribution score of each feature is better.
6. In comparison to LIME, SHAP took longer to train on the dataset.
7. The LDA model was applied on the positive and negative customer review dataset, and the reviews were split into various topics.
8. The optimum model was picked on basis of high coherence score with number of topics.
9. The topics along with its keywords can be used by FDS organisations for identifying problems and solve them.

5 CONCLUSIONS AND FUTURE WORK RECOMMENDATIONS

5.1 General Conclusion

This research aimed to predict the sentiment from customer reviews in the FDS domain and explain the predictions. AI can assist FDS organisations in solving problems and saving money, given the large amount of review data spread across multiple platforms and the lack of customer service consultants to go through and act on each of these comments. In the FDS domain, false positive indicates more operational efforts, whereas false negative increases the risk for an organisation to miss important customer complaints. The results showed that the LSTM model with lower false negatives outperforms the BiLSTM and Bi-GRU-LSTM-CNN models. SHAP and LIME were successfully applied to the LSTM model for determining the positive or negative contributions of each word on the predictions made by the model. Original customer reviews were analysed, and understanding the logic behind the predictions made by the DL models, such as LSTM, was possible. Therefore, this research revealed that the behaviour of the models can be discovered by implementing DL models for sentiment analysis along with XAI techniques. LIME explainer explains what features contribute to particular prediction, and SHAP explainer can further deepen the understanding. SHAP takes more time in training with the dataset compared with LIME. This research concludes that the sentiment analysis of customer reviews in FDS can be best achieved with the LSTM model combined with LIME and SHAP techniques for achieving high accuracy and explainability. Further, with the help of LDA model various topics were identified from negative and positive sentiment dataset.

5.2 Conclusion of Objective 1

In this study, three DL models (LSTM, Bi-LSTM, BiGRU-LSTM-CNN) are developed on customer review dataset by fine tunning the hyper parameters after multiple rounds of training and testing. LSTM, Bi-LSTM and Bi-GRU-LSTM-CNN obtained an accuracy of 96.07%, 95.85% and 96.33%, respectively. The FDS organisations aim to identify and address each and every customer complaint without missing any of them to improve

customer satisfaction. Thus, the model's prediction should have fewer false negatives. Given that all the DL models achieved close accuracy levels, the model with lesser false negatives was selected. Although, the accuracy of Bi-GRU-LSTM-CNN is 96.33%, its false negative percentage is 2.13, which is more than the LSTM model (0.77). The overall accuracy of the LSTM model is 96.07%. In the case of sentiment analysis, lesser rate in false negative is preferred over false positive because businesses do not like to miss any negative customer reviews as compared to positive. The LSTM model is recommended over the Bi-LSTM and Bi-GRU-LSTM-CNN models due to its lower false negative percentage. This model can be used in performing sentiment analysis on customer reviews for any FDS organisation. The main findings of Objective 1 are as follows:

- Three different DL models (LSTM, Bi-LSTM, Bi-GRU-LSTM-CNN) were trained and were able to perform sentiment analysis on customer reviews.
- F1 Score of all the three DL models came as 97 for detecting customer complaints.
- Recall of LSTM model is better than Bi-LSTM and Bi-GRU-LSTM-CNN model.
- LSTM model came up with low false negative value for detecting review as complaints as compared to Bi-LSTM and Bi-GRU-LSTM-CNN model.
- LSTM is picked as the best DL model for performing sentiment analysis in FDS domain.

5.3 Conclusion of Objective 2

In this study, SHAP and LIME techniques are used to interpret the feature importance considered by LSTM model when making predictions on the customer reviews. SHAP and LIME allowed us to perform an in-depth analysis of the model with its sample customer review test data. SHAP and LIME successfully verified the prediction of the LSTM model by looking at the features which were contributing the negative and positive outcome. SHAP's ability to show the interpretation of LSTM predictions by pinpointing the contribution score of each feature is better as compared to that of LIME. However, SHAP took more time to train with the dataset compared with LIME. Hence, using SHAP and LIME explanations, FDS organisations can use DL model for performing sentiment analysis. The main findings of the Objective 2 are as follows:

- SHAP and LIME XAI method applied successfully on LSTM model to determine the contributions made by the words in predicting the outcome.
- SHAP's ability to show the interpretation of LSTM predictions by pinpointing the contribution score of each feature is better as compared to that of LIME.
- SHAP took more time to train with the dataset compared with LIME.
- Hence SHAP and LIME method can be used to interpret the DL models in FDS domain to verify the prediction logic.

5.4 Conclusion of Objective 3

In this study research shows that using LDA modelling technique, various topics were identified from positive and negative customer review dataset. The optimum number of topics which can be extracted from the dataset, can be calculated based on higher coherence score. The research found the LDA model for negative reviews had 11 topics selected on the basis of higher coherence value. Similarly, the LDA model for positive reviews had 3 topics. The study found the negative comments can be grouped into various topics such as "Account", "Delivery", "Order", "Food Quality" and "Online customer service". The concerned department for these topics can look into keywords for further details on which the improvements can be made. Similarly, the topics such as "Food Quality" and "Customer Service" coming from positive dataset can be used for awarding staffs and restaurants. The main findings of Objective 3 are as follows:

- The LDA model was applied on the positive and negative customer review dataset, and the reviews were split into various topics.
- The optimum model was picked on basis of high coherence score with number of topics.
- The topics along with its keywords can be used by FDS organisations for identifying problems and solve them.
- Other topic modelling techniques can be implemented to overcome to issue of overlapping small bubbles as observed in the interactive chart.

5.5 Research Drawbacks and Limitations

The research implemented DL models for performing sentiment analysis on customer reviews on FDS domain. In presence of sarcasm in review dataset, DL models can misinterpret whether the customer review is good one or vice versa, thereby resulting in an improper training set. Also, the risk of spam accounts, false accounts and bots, can generate irrelevant data and affect the training set. These issues can affect the accuracy of the DL models. The research found in case of topic modelling using LDA on negative customer reviews, many overlaps and small sized bubbles clustered were formed in one region which suggested many topics in those areas. Those topics could not be properly identified as different keywords were pointing to different topic group.

5.6 Recommendations for Future Work

In this research, the proposed models and methods were implemented and all the three objectives were achieved. Moreover, further work can be done on sentiment analysis in FDS domain by applying the latest models such as Bidirectional Encoder Representations from Transformers (BERT). The accuracy score of BERT can be compared with the accuracy of the current DL model's outcome. Furthermore, more research can be done to implement new methods of XAI technique which can interpret the model better as compared to SHAP and LIME. More, topic modelling techniques such as Latent Semantic Analysis can be explored to get better results out of the dataset. The model and the framework developed in this research can be used on multiple FDS platforms so that DL models get trained with larger and variety of dataset. In this research we have used Productreview website to gather data but that can be mixed with other FDS platforms to see the behaviour. This will add more testing with data coming from different platforms. Also, engaging with stakeholders would be nice idea to deploy the solution. It will benefit FDS organisation to get more insights into customer needs. Also, the combination of the DL model along with XAI techniques and LDA model can be used in other domains to look into the outcome. Finally, it will be worth to check the model working in real time scenario where as soon as customer lodges complaint, the complaint goes to its respective department for its resolution.

REFERENCES

Abdelwahab, Ahmed, Jose Robles, Costin-Gabriel Chiru, and Traian Rebedea. 2014. 'Tweets Topic Modelling Across Different Countries', *eLearning & Software for Education*.

Ajit, Arohan, Koustav Acharya, and Abhishek Samanta. 2020. "A review of convolutional neural networks." In *2020 international conference on emerging trends in information technology and engineering (ic-ETITE)*, 1-5. IEEE.

Akila, R., S. Revathi, and G. Shreedevi. 2020. "Opinion Mining on Food Services using Topic Modeling and Machine Learning Algorithms." In *2020 6th International Conference on Advanced Computing and Communication Systems, ICACCS 2020*, 1071-76.

Ali, Farman, Daehan Kwak, Pervez Khan, Shaker El-Sappagh, Amjad Ali, Sana Ullah, Kye Hyun Kim, and Kyung-Sup Kwak. 2019. 'Transportation sentiment analysis using word embedding and ontology-based topic modeling', *Knowledge-based systems*, 174: 27-42.

Alshari, Eissa M, Azreen Azman, Shyamala Doraisamy, Norwati Mustapha, and Mostafa Alkeshr. 2018. "Effective method for sentiment lexical dictionary enrichment based on Word2Vec for sentiment analysis." In *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 1-5. IEEE.

Ancona, Marco, Cengiz Öztireli, and Markus Gross. 2019. 'Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Values Approximation'.

Ara, J., M. T. Hasan, A. Al Omar, and H. Bhuiyan. 2020. "Understanding Customer Sentiment: Lexical Analysis of Restaurant Reviews." In *2020 IEEE Region 10 Symposium, TENSYMP 2020*, 295-99.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58: 82-115.

Barsky, Jonathan D., and Richard Labagh. 1992. 'A Strategy for Customer Satisfaction', *The Cornell hotel and restaurant administration quarterly*, 33: 32-40.

Berger, Adam, Stephen A Della Pietra, and Vincent J Della Pietra. 1996. 'A maximum entropy approach to natural language processing', *Computational linguistics*, 22: 39-71.

Bhuiyan, M. R., M. H. Mahedi, N. Hossain, Z. N. Tumpa, and S. A. Hossain. 2020. "An Attention Based Approach for Sentiment Analysis of Food Review Dataset." In *2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020*.

Bittermann, André, and Andreas Fischer. 2018. 'How to identify hot topics in psychology using topic modeling', *Zeitschrift für Psychologie*.

Brandt, Tobias, Johannes Bendler, and Dirk Neumann. 2017. 'Social media analytics and value creation in urban smart tourism ecosystems', *Information & Management*, 54: 703-13.

Calheiros, Ana Catarina, Sérgio Moro, and Paulo Rita. 2017. 'Sentiment classification of consumer-generated online reviews using topic modeling', *Journal of Hospitality Marketing & Management*, 26: 675-93.

Cambray, Aleix, and Norbert Podsadowski. 2019. 'Bidirectional recurrent models for offensive tweet classification', *arXiv preprint arXiv:1903.08808*.

Chandriah, Kiran Kumar, and Raghavendra V Naraganahalli. 2021. 'RNN/LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting', *Multimedia Tools and Applications*, 80: 26145-59.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 'Learning phrase representations using RNN encoder-decoder for statistical machine translation', *arXiv preprint arXiv:1406.1078*.

Dang, Nhan Cach, María N Moreno-García, and Fernando De la Prieta. 2020. 'Sentiment analysis based on deep learning: A comparative study', *Electronics*, 9: 483.

Di Cicco, Vincenzo, Donatella Firmani, Nick Koudas, Paolo Merialdo, and Divesh Srivastava. 2019. "Interpreting deep learning models for entity resolution: an experience report using LIME." In *International Conference on Management of Data*, 1-4. ACM.

Domingos, Pedro, and Michael Pazzani. 1997. 'On the optimality of the simple Bayesian classifier under zero-one loss', *Machine learning*, 29: 103-30.

Drus, Zulfadzli, and Haliyana Khalid. 2019. 'Sentiment analysis in social media and its application: Systematic literature review', *Procedia Computer Science*, 161: 707-14.

Failory.com. 2017. 'What was Sprig?', Accessed July 5. https://www.failory.com/cemetery/sprig.

Fellbaum, Christiane. 2017. '16 WordNet: An Electronic Lexical Resource', *The Oxford handbook of cognitive science*: 301.

Geler, Z., M. Savić, B. Bratić, V. Kurbalija, M. Ivanović, and W. Dai. 2021. 'Sentiment prediction based on analysis of customers assessments in food serving businesses', *Connection Science*.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. 'A survey of methods for explaining black box models', *ACM computing surveys (CSUR)*, 51: 1-42.

Hasan, Mahedi, Anichur Rahman, Md Karim, Md Khan, Saikat Islam, and Md Islam. 2021. "Normalized approach to find optimal number of topics in Latent Dirichlet Allocation (LDA)." In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*, 341-54. Springer.

He, W., X. Tian, R. Tao, W. Zhang, G. Yan, and V. Akula. 2017. 'Application of social media analytics: A case of analyzing online hotel reviews', *Online Information Review*, 41: 921-35.

Hegde, S. B., S. Satyappanavar, and S. Setty. 2018. "Sentiment based Food Classification for Restaurant Business." In *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, 1455-62.

Hung, B. T. 2020. "Integrating Sentiment Analysis in Recommender Systems." In *Springer Series in Reliability Engineering*, 127-37.

Jelodar, Hamed, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. 'Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey', *Multimedia Tools and Applications*, 78: 15169-211.

Jeong, Byeongki, Janghyeok Yoon, and Jae-Min Lee. 2019. 'Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis', *International Journal of Information Management*, 48: 280-90.

Jiang, Y. 2020. "Restaurant reviews analysis model based on machine learning algorithms." In *Proceedings - 2020 Management Science Informatization and Economic Innovation Development Conference, MSIEID 2020*, 169-78.

Joachims, Thorsten. 1998. "Text categorization with support vector machines: Learning with many relevant features." In *European conference on machine learning*, 137-42. Springer.

Johnson, Rie, and Tong Zhang. 2014. 'Effective use of word order for text categorization with convolutional neural networks', *arXiv preprint arXiv:1412.1058*.

Kenny, E. M., E. D. Delaney, D. Greene, and M. T. Keane. 2021. "Post-hoc Explanation Options for XAI in Deep Learning: The Insight Centre for Data Analytics Perspective." In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 20-34.

Kenny, E. M., C. Ford, M. Quinn, and M. T. Keane. 2021. 'Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies', *Artificial Intelligence*, 294.

Kim, Been, Oluwasanmi Koyejo, and Rajiv Khanna. 2016. "Examples are not enough, learn to criticize! Criticism for Interpretability." In *NIPS*, 2280-88.

Kim, Sung Guen, and Juyoung Kang. 2018. 'Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews', *Information Processing & Management*, 54: 938-57.

Kim, Yoon. 2014. 'Convolutional Neural Networks for Sentence Classification'.

Kitchenham, Barbara, and Stuart Charters. 2007. 'Guidelines for performing systematic literature reviews in software engineering'.

Krishnakumari, K, E Sivasankar, and Sam Radhakrishnan. 2020. 'Hyperparameter tuning in convolutional neural networks for domain adaptation in sentiment classification (HTCNN-DASC)', *Soft Computing*, 24: 3511-27.

Krouska, Akrivi, Christos Troussas, and Maria Virvou. 2020. 'Deep learning for twitter sentiment analysis: the effect of pre-trained word embedding.' in, *Machine learning paradigms* (Springer).

Laguna, Laura, Susana Fiszman, Patricia Puerta, C Chaya, and Amparo Tárrega. 2020. 'The impact of COVID-19 lockdown on food priorities. Results from a preliminary study using social media and an online survey with Spanish consumers', *Food Quality and Preference*, 86: 104028.

Lan, Hong, LI Ya'nan, and Wang Shuhua. 2016. 'Improvement of online food delivery service based on consumers' negative comments', *Canadian Social Science*, 12: 84-88.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. 'Deep learning', *nature*, 521: 436-44.

Lewis, David D. 1998. "Naive (Bayes) at forty: The independence assumption in information retrieval." In *European conference on machine learning*, 4-15. Springer.

Liz, H., M. Sánchez-Montañés, A. Tagarro, S. Domínguez-Rodríguez, R. Dagan, and D. Camacho. 2021. 'Ensembles of Convolutional Neural Network models for pediatric pneumonia diagnosis', *Future Generation Computer Systems*, 122: 220-33.

Lokeshkumar, R., O. V. Sabnis, and S. Bhattacharyya. 2020. "A Novel Approach to Extract and Analyse Trending Cuisines on Social Media." In *Lecture Notes on Data Engineering and Communications Technologies*, 645-56.

Lopez, Marc Moreno, and Jugal Kalita. 2017. 'Deep Learning applied to NLP', *arXiv preprint arXiv:1703.03091*.

Lorente, M. P. S., E. M. Lopez, L. A. Florez, A. L. Espino, J. A. I. Martínez, and A. S. de Miguel. 2021. 'Explaining deep learning-based driver models', *Applied Sciences (Switzerland)*, 11.

Lundberg, Scott, and Su-In Lee. 2017. 'A unified approach to interpreting model predictions', *arXiv preprint arXiv:1705.07874*.

Luo, Y., L. Tang, E. Kim, and X. Wang. 2020. 'Finding the reviews on yelp that actually matter to me: Innovative approach of improving recommender systems', *International Journal of Hospitality Management*, 91.

Luo, Y., and X. Xu. 2019. 'Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp', *Sustainability (Switzerland)*, 11.

Luo, Y., and X. Xu. 2021. 2021. 'Comparative study of deep learning models for analyzing online restaurant reviews in the era of the COVID-19 pandemic', *International Journal of Hospitality Management*, 94.

Mabrouk, Alhassan, Rebeca P Díaz Redondo, and Mohammed Kayed. 2020. 'Deep learning-based sentiment classification: A comparative survey', *IEEE Access*, 8: 85616-38.

Mangal, Sanidhya, Poorva Joshi, and Rahul Modak. 2019. 'LSTM vs. GRU vs. Bidirectional RNN for script generation', *arXiv preprint arXiv:1908.04332*.

Mathayomchan, B., and V. Taecharungroj. 2020. "“How was your meal?” Examining customer experience using Google maps reviews', *International Journal of Hospitality Management*, 90.

Mathews, S. M. 2019. "Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review." In *Advances in Intelligent Systems and Computing*, 1269-92.

Mhlanga, Oswald. 2018. 'The fast food industry in South Africa: the micro-environment and its influence', *African Journal of Hospitality, Tourism and Leisure*.

Molnar, Christoph. 2020. *Interpretable machine learning* (Lulu. com).

Moradi, M., and M. Samwald. 2021. 'Post-hoc explanation of black-box classifiers using confident itemsets', *Expert Systems with Applications*, 165.

Moraes, Rodrigo, João Francisco Valiati, and Wilson P GaviãO Neto. 2013. 'Document-level sentiment classification: An empirical comparison between SVM and ANN', *Expert Systems with Applications*, 40: 621-33.

Muhammad, B. A., R. Iqbal, A. James, and D. Nkantah. 2020. "Comparative Performance of Machine Learning Methods for Text Classification." In *2020 International Conference on Computing and Information Technology, ICCIT 2020*.

Nagpal, M., K. Kansal, A. Chopra, N. Gautam, and V. K. Jain. 2020. "Effective Approach for Sentiment Analysis of Food Delivery Apps." In *Advances in Intelligent Systems and Computing*, 527-36.

Nigam, Kamal, John Lafferty, and Andrew McCallum. 1999. "Using maximum entropy for text classification." In *IJCAI-99 workshop on machine learning for information filtering*, 61-67. Stockholom, Sweden.

Panda, Geetanjali, Ashwani Kumar Upadhyay, and Komal Khandelwal. 2019. 'Artificial intelligence: A strategic disruption in public relations', *Journal of Creative Communications*, 14: 196-213.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. 'Thumbs up? Sentiment classification using machine learning techniques', *arXiv preprint cs/0205070*.

Parliament of Australia. 2018. 'Population and migration statistics in Australia', Accessed July 5. https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp1819/Quick_Guides/PopulationStatistics.

Poelman, Maartje P, Marleen Gillebaart, Caroline Schlinkert, S Coosje Dijkstra, Elianne Derksen, Frederike Mensink, Roel CJ Hermans, Pleun Aardening, Denise de Ridder, and Emely de Vet. 2021. 'Eating behavior and food purchases during the COVID-19 lockdown: A cross-sectional study among adults in the Netherlands', *Appetite*, 157: 105002.

Psychoula, Ismini, Andreas Gutmann, Pradip Mainali, Sharon H Lee, Paul Dunphy, and Fabien Petitcolas. 2021. 'Explainable machine learning for fraud detection', *Computer*, 54: 49-59.

Ravi, Kumar, and Vadlamani Ravi. 2015. 'A survey on opinion mining and sentiment analysis: tasks, approaches and applications', *Knowledge-based systems*, 89: 14-46.

Reiley, Laura. 2020. 'A pandemic surge in food delivery has made ghost kitchens and virtual eateries one of the only growth areas in the restaurant industry', *The Washington post*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "" Why should i trust you?" Explaining the predictions of any classifier." In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135-44.

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. "Exploring the space of topic coherence measures." In *Proceedings of the eighth ACM international conference on Web search and data mining*, 399-408.

Samek, W., G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller. 2021. 'Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications', *Proceedings of the IEEE*, 109: 247-78.

Schmidhuber, Jürgen, and Sepp Hochreiter. 1997. 'Long short-term memory', *Neural Comput*, 9: 1735-80.

Schmiedel, Theresa, Oliver Müller, and Jan vom Brocke. 2019. 'Topic modeling as a strategy of inquiry in organizational research: A tutorial with an application example on organizational culture', *Organizational Research Methods*, 22: 941-68.

Schoenborn, J. M., and K. D. Althoff. 2019. "Recent trends in XAI: A broad overview on current approaches, methodologies and interactions." In *CEUR Workshop Proceedings*, 51-60.

Scholkopf, B. 1999. 'Making large scale SVM learning practical', *Advances in Kernel Methods: Support Vector Learning*: 41-56.

Shaeeali, Noor Sakinah, Azlinah Mohamed, and Sofianita Mutalib. 2020. 'Customer reviews analytics on food delivery services in social media: a review', *IAES International Journal of Artificial Intelligence*, 9: 691.

Shankaranarayana, S. M., and D. Runje. 2019. "ALIME: Autoencoder based approach for local interpretability." In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 454-63.

Sharif, O., M. M. Hoque, and E. Hossain. 2019. "Sentiment Analysis of Bengali Texts on Online Restaurant Reviews Using Multinomial Naïve Bayes." In *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*.

Singh, A., S. Sengupta, and V. Lakshminarayanan. 2020. 'Explainable deep learning models in medical image analysis', *Journal of Imaging*, 6.

Singh, R. K., and H. K. Verma. 2020. 'Influence of Social Media Analytics on Online Food Delivery Systems', *International Journal of Information System Modeling and Design*, 11: 1-21.

So, Chaehan. 2020. "Understanding the Prediction Mechanism of Sentiments by XAI Visualization." In *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 75-80.

Statista. 2021. 'Online Food Delivery', Accessed July 5. https://www.statista.com/outlook/dmo/eservices/online-food-delivery/australia.

Suciati, A., and I. Budi. 2019. "Aspect-based Opinion Mining for Code-Mixed Restaurant Reviews in Indonesia." In *Proceedings of the 2019 International Conference on Asian Language Processing, IALP 2019*, 59-64.

Sue, Mitchell. 2018. 'Menulog moves to add delivery services', *The Australian financial review*.

Suhartanto, Dwi, Mohd Helmi Ali, Kim Hua Tan, Fauziyah Sjahroeddin, and Lusianus Kusdibyo. 2019. 'Loyalty toward online food delivery service: the role of e-service quality and food quality', *Journal of foodservice business research*, 22: 81-97.

Tang, Duyu, Bing Qin, and Ting Liu. 2015. "Document modeling with gated recurrent neural network for sentiment classification." In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1422-32.

Techcrunch.com. 2019. 'After raising $125M, Munchery fails to deliver', Accessed July 5. https://techcrunch.com/2019/01/21/munchery-shuts-down/.

Thikshaja, Uthra Kunathur, and Anand Paul. 2018. 'A brief review on deep learning and types of implementation for deep learning', *Deep Learning Innovations and Their Convergence With Big Data*: 20-32.

Tian, G., L. Lu, and C. McIntosh. 2021. 'What factors affect consumers' dining sentiments and their ratings: Evidence from restaurant online review data', *Food Quality and Preference*, 88.

Upadhyay, Anand, Swapnil Rai, and Sneha Shukla. 2022. "Sentiment Analysis of Zomato and Swiggy Food Delivery Management System." In *Second International Conference on Sustainable Technologies for Computational Intelligence*, edited by Ashish Kumar Luhach, Ramesh Chandra Poonia, Xiao-Zhi Gao and Dharm Singh Jat, 39-46. Singapore: Springer Singapore.

Utkin, Lev V, Anna A Meldo, Maxim S Kovalev, and Ernest M Kasimov. 2020. "A Simple General Algorithm for the Diagnosis Explanation of Computer-Aided Diagnosis Systems in Terms of Natural Language Primitives." In *2020 XXIII International Conference on Soft Computing and Measurements (SCM)*, 202-05. IEEE.

Westerlund, Mika, Zarin Mahmood, Seppo Leminen, and Mervi Rajahonka. 2019. "Topic modelling analysis of online reviews: Indian restaurants at Amazon. Com." In *ISPIM Conference Proceedings*, 1-14. The International Society for Professional Innovation Management (ISPIM).

Widodo Wijayanto, U., and R. Sarno. 2018. "An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naïve Bayes." In *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018*, 476-81.

Windasari, Ike Pertiwi, and Dania Eridani. 2017. "Sentiment analysis on travel destination in Indonesia." In *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 276-79. IEEE.

Wolanin, A., G. Mateo-Garciá, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi, and L. Guanter. 2020. 'Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt', *Environmental Research Letters*, 15.

Xiao, Min, and Yuhong Guo. 2015. "Annotation projection-based representation learning for cross-lingual dependency parsing." In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, 73-82.

Yin, Wenpeng, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. 'Comparative study of CNN and RNN for natural language processing', *arXiv preprint arXiv:1702.01923*.

Yu, C. E., and X. Zhang. 2020. 'The embedded feelings in local gastronomy: a sentiment analysis of online reviews', *Journal of Hospitality and Tourism Technology*, 11: 461-78.

Zachary Kefa, Chepukaka, and Kirugi Fridah Kendi. 2019. 'Service Quality And Customer Satisfaction at Kenya National Archives and Documentation Service, Nairobi County: Servqual Model Revisited', *International Journal on Customer Relations*, 7: 1.

Zahoor, K., N. Z. Bawany, and S. Hamid. 2020. "Sentiment analysis and classification of restaurant reviews using machine learning." In *Proceedings - 2020 21st International Arab Conference on Information Technology, ACIT 2020*.

Zhang, Lei, Shuai Wang, and Bing Liu. 2018. 'Deep learning for sentiment analysis: A survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8: e1253.

Zhang, Xiang, Junbo Zhao, and Yann LeCun. 2015. 'Character-level convolutional networks for text classification', *Advances in neural information processing systems*, 28.

Zucco, Chiara, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro. 2018. "Explainable sentiment analysis with applications in medicine." In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1740-47. IEEE.