LARGE LANGUAGE MODEL-BASED CUSTOMER FEEDBACK ANALYSIS IN HOSPITALITY DOMAIN FOR IMPROVEMENT SUGGESTIONS

SENDURAN R

239357C

Masters in Computer Science

Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2025

LARGE LANGUAGE MODEL-BASED CUSTOMER FEEDBACK ANALYSIS IN HOSPITALITY DOMAIN FOR IMPROVEMENT SUGGESTIONS

SENDURAN R

239357C

Thesis submitted in partial fulfillment of the requirements for the degree
Masters in Computer Science

Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

June 2025

DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 23/06/2025

The supervisor should certify the Thesis with the following declaration.

The above candidate has carried out research for the Masters in Computer Science Thesis under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Indika Perera

Signature of the Supervisor: Date: 25/06/2025

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to everyone who has supported me throughout the journey of completing this research. First and foremost, I extend my sincere thanks to my supervisor, Dr Indika Perera, for their invaluable guidance, encouragement, and constructive feedback, which have been instrumental in shaping this work. Your expertise and insights have been a beacon of inspiration throughout this journey. I am also deeply grateful to Faculty of Computer Science and Engineering, University of Moratuwa for providing the necessary resources and a conducive environment to conduct this research. I would like to acknowledge the developers and contributors behind the Large Language Models (LLMs) and tools that played a crucial role in the implementation of this study. Their work has laid the foundation for innovative research like mine. This research would not have been possible without the collective efforts and inspiration of all these individuals and institutions.

ABSTRACT

This research investigates the challenge of systematically transforming customer feedback into actionable service improvement suggestions in the hospitality sector, where unstructured reviews often go underutilized. To address this, we propose a novel three-stage natural language processing (NLP) pipeline that utilizes fine-tuned large language models (LLMs) designed for efficiency, scalability, and domain adaptation. The study uses a dataset of approximately 4,300 recent reviews collected from TripAdvisor and Booking.com to train and evaluate each stage of the pipeline. First, a sentiment classification model (DistilBERT-base-uncased) categorizes reviews as fully positive or negative. Next, a fine-tuned sequence-to-sequence model (Flan-T5-base) extracts key negative aspects. Finally, an open-ended suggestion generation stage is performed using a fine-tuned LLM (Mistral-7B-Instruct). Unlike prior approaches that focus narrowly on sentiment or rely on manual rule engineering, our pipeline offers a fully automated, end-to-end framework tailored for high-volume review environments. The system is evaluated using standard metrics. The sentiment classification stage achieved over 92% accuracy with high precision, recall, and F1-scores. The aspect extraction stage attained a BLEU score of 0.65 and ROUGE-L of 0.85. The final suggestion generation stage produced coherent outputs. Although the BLEU score was low due to mismatch with references, human evaluation confirmed the suggestions were contextually appropriate. This study contributes to applied NLP by demonstrating how lightweight, fine-tuned models can be integrated to produce coherent, domain-specific, and scalable insights for real-world service improvement. The approach opens avenues for further research in multi-domain adaptation, real-time feedback analytics, and explainable AI for recommendation systems.

Keywords: Natural Language Processing, Sentiment Analysis, Aspect Extraction, Text Generation, Customer Feedback, Hospitality Industry, Service Improvement, DistilBERT, Flan-T5, Mistral-7B-Instruct, Fine-tuning, Large Language Models, Review Analysis, NLP Pipeline, Scalable AI Systems

TABLE OF CONTENTS

Declaration of the Candidate & Supervisor i

Acknowledgement ii

Abstract iii

Table of Contents iv

List of Figures viii

List of Tables x

List of Abbreviations x

1 Introduction 1

1.1 Motivation 1

1.1.1 Customer-Centricity and Brand Reputation 2

1.1.2 Navigating Rapid Market Changes 2

1.1.3 Leveraging Technological Advancements 3

1.1.4 Rise of Automated Feedback Analysis 4

1.1.5 Emergence of Large Language Models (LLMs) 4

1.1.6 Advancing Beyond Traditional NLP Methods using LLMs 5

1.2 Research Problem and Objectives 7

1.2.1 Research Problem 7

1.2.2 Objectives 10

2 Literature Review 11

2.1 LLM - Transformer Architectures 11

2.1.1 What is a Transformer? 11

2.1.2 Attention Mechanism 11

2.1.3 Encoder-only Transformers (BERT, DistilBERT) 12

2.1.4 Encoder-Decoder Transformers (T5, Flan-T5, BART) 12

2.1.5 Decoder-only Transformers (GPT, Mistral, LLaMA) 12

2.1.6 Comparison of Architectures 13

2.2 Relevant researches using Large Language Models 13

2.3 Relevant researches using Machine Learning Algorithms 19
2.4 Baseline Models and Comparison 20
2.4.1 BERT vs RoBERTa vs DistilBERT for sentiment classification 21
2.4.2 GPT variants or T5-small for negative comment extraction 23
2.4.3 Non-LLM approaches: SVM, logistic regression, or rule-based negativity extraction 25
2.4.4 Comparison Criteria 26
3 Methodology 28
3.1 Overview of Methodology 28
3.2 Levels of Using LLMs 28
3.3 Prompt Engineering 29
3.4 LLM Fine-tuning 31
3.5 Supervised Fine-tuning 32
3.6 Parameter Efficient Fine-tuning (PEFT) 33
3.7 Research Design 34
3.7.1 Proposed Pipeline 34
3.7.2 Overall Research Framework 35
3.7.3 Rationale for a Three-Step Pipeline 36
3.8 Perspective on Models Selection 36
3.8.1 Research Questions and Hypotheses 37
3.9 Data Collection and Preparation 38
3.9.1 Data Sources 38
3.9.2 Distillation 39
3.9.3 Data Distillation 40
3.9.4 Application of Data Distillation in Service Improvement Suggestion Models building 40
3.9.5 Data Annotation 42
3.9.6 Data Cleaning and Pre-processing 43
3.9.7 Data Augmentation 45
3.9.8 Why Data Augmentation was not necessary! 46
3.10 Preliminary Experiments 47

3.10.1 distilbert-base-uncased for Sentiment Analysis 47

3.10.2 distilbert-base-uncased-finetuned-sst-2-english for Sentiment Analysis 47

3.10.3 Baseline Attempts and Prompt Engineering with google/Flan-T5-base 47

3.10.4 mistralai/Mistral-7B-Instruct-v0.1 outputs for negative aspect extraction before fine-tuning using our dataset 51

3.10.5 mistralai/Mistral-7B-Instruct-v0.1 outputs for service improvement suggestion before fine-tuning using our dataset 52

4 Implementation 54

4.1 Implementation Steps 54

4.1.1 Dataset Collection and Preparation 55

4.1.2 Step 1 - Sentiment Analysis using distilBERT-base-uncased 58

4.1.3 Step 2 - Extracting negative aspects using Flan-T5 model 62

4.1.4 Step 3 - Service Improvement Suggestion Generation using fine-tuned mistralai/Mistral-7B-Instruct-v0.1 65

4.2 Challenges in Implementation 67

4.3 Tools Used 69

5 Evaluation and Result 70

5.1 Overview of Result 70

5.2 Accuracy, Precision, Recall, and F1-Score 70

5.2.1 Accuracy 70

5.2.2 Precision 71

5.2.3 Recall 71

5.2.4 F1-Score 71

5.2.5 Confusion Matrix 72

5.2.6 ROC Curve & AUC Score 72

5.2.7 PR Curve & AUC Score 72

5.3 BLEU and ROUGE for Text Generation 72

5.3.1 BLEU (Bilingual Evaluation Understudy) 72

5.3.2 ROUGE (Recall-Oriented Understudy for Gisting Evaluation) 73

5.4 DistilBERT-base-uncased Sentiment Analysis: Evaluation and Results 73
5.5 Google/flan-t5-base Negative Aspects Extraction: Evaluation and Results 77
5.6 Mistralai/Mistral-7B-Instruct-v0.1 Service Improvment Suggestion Generation: Evaluation and Results 81
6 Conclusion 83
6.1 Summary of the Three-Step Approach 83
6.2 Room for Improvement 84
6.3 Research Contributions 85
6.3.1 Research Limitations 86
6.4 Learning outcome of this research 87
6.5 Future Work 87
6.5.1 Applying on Various Domains 88
6.5.2 Utilizing different data sources 88
6.5.3 Incorporating Retrieval-Augmented Generation (RAG) 88
6.5.4 Improving Model Robustness and Interpretability 88
6.5.5 User Feedback Integration and Continual Learning 89
6.5.6 Deployment and Real-world Validation 89
References 90

LIST OF FIGURES

Figure	Description	Page
Figure 1.1	A Sample Hotel Review 1	8
Figure 1.2	A Sample Hotel Review 2	8
Figure 1.3	A Sample Hotel Review 3	9
Figure 3.1	Flow Digram - Overall Methodology	34
Figure 3.2	Flow Digram - Combined Negative Aspects for Generating Service Improvement Suggestions	36
Figure 3.3	Review (distilbert-base-uncased output)	48
Figure 3.4	Review (distilbert-base-uncased-finetuned-sst-2-english output)	49
Figure 3.5	Review (Expected Sentiment Output with Reasons)	50
Figure 3.6	Review (distilbert-base-uncased output)	51
Figure 3.7	Sample Review tested on flan-t5-based non fine-tuned	51
Figure 3.8	mistralai/Mistral-7B-Instruct-v0.1 outputs for negative aspect extraction before fine-tuning using our dataset	52
Figure 3.9	mistralai/Mistral-7B-Instruct-v0.1 outputs for service improvement suggestion before fine-tuning using our dataset	53
Figure 4.1	Flow Digram	54
Figure 4.2	Booking.com Response 1	55
Figure 4.3	Booking.com Response 2	55
Figure 4.4	Triadvisor Response 1	56
Figure 4.5	Triadvisor Response 2	56
Figure 4.6	Sentiment Analysis Dataset for Reviews	58
Figure 4.7	Distribution of Sentiment Labels of Reviews	59
Figure 4.8	Distribution of word counts	59
Figure 4.9	Distillbert Overall Architecture	61
Figure 4.10	Generated Output of google/flan-t5-base after fine tuned	63
Figure 4.11	Flan-t5-base Overall Architecture	64
Figure 4.12	Review on Post-Training Evaluation and Inference of flan-t5-base	65
Figure 4.13	Result on Post-Training Evaluation and Inference of flan-t5-base	65
Figure 4.14	mistralai/Mistral-7B-Instruct-v0.1 Overall Architecture	66
Figure 4.15	fine-tuned mistralai/Mistral-7B-Instruct-v0.1 output for Service Improvement Suggestion Generation	68
Figure 5.1	Confusion Matrix - Sentiment Analysis	76
Figure 5.2	Precision-Recall Curve	76
Figure 5.3	ROC Curve	77

Figure 5.4	BLEU-n Precision Comparison: Full vs LoRA based fine-tuning	80
Figure 5.5	ROUGE Score Comparison: Full vs LoRA based fine-tuning	80

LIST OF TABLES

Table	Description	Page
Table 2.1	Comparison of Transformer Architectures Used in This Research	13
Table 4.1	FLAN-T5 Family and Trainable Parameters	62
Table 5.1	DistilBERT-base-uncased : Training and Validation Loss per Epoch	74
Table 5.2	DistilBERT-base-uncased : Accuracy and Negative Class Metrics per Epoch	74
Table 5.3	DistilBERT-base-uncased : Overall Training Summary	75
Table 5.4	google/flan-t5-base : BLEU Score Breakdown for Negative Aspect Extraction on fully fine-tuned	78
Table 5.5	google/flan-t5-base : ROUGE Score Breakdown for Negative Aspect Extraction on fully fine-tuned	78
Table 5.6	google/flan-t5-base : BLEU Score Breakdown for Negative Aspect Extraction on LoRA based fine-tuned	79
Table 5.7	google/flan-t5-base : ROUGE Score Breakdown for Negative Aspect Extraction on LoRA based fine-tuned	79
Table 5.8	mistralai/Mistral-7B-Instruct-v0.1 : BLEU Score Breakdown for Service Improvement Suggestion Generation	81
Table 5.9	mistralai/Mistral-7B-Instruct-v0.1 : ROUGE Score Breakdown for Service Improvement Suggestion Generation	82

LIST OF ABBREVIATIONS

Abbreviation	**Description**
AUC	Area Under the Curve
BLEU	Bilingual Evaluation Understudy
CNNs	Convolutional Neural Networks
LLMs	Large Language Models
LoRA	Low-Rank Adaptation
NLP	Natural Language Processing
PEFT	Parameter Efficient Fine-tuning
PR	Precision-Recall
RAG	Retrieval-Augmented Generation
RNNs	Recursive Neural Networks
ROC	Receiver Operating Characteristic
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SVM	Support Vector Machine