AN ATTENTION-BASED RECURRENT NEURAL NETWORK MODEL FOR IMPROVED ASPECT-BASED SENTIMENT ANALYSIS OF CUSTOMER REVIEWS

Kevin Nyakundi Onsale

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Science of Murang'a University of Technology

August 2025

DECLARATION

Student Declaration

I hereby declare that this Thesis is my original work and to the best of my knowledge has not been presented for a degree award in this or any other university. I understand what plagiarism is and I am aware of the University's policy in this regard. I understand that any false claim in respect of this work shall result in disciplinary action in accordance with the University Anti-Plagiarism Policy.

_____ _____

Kevin Nyakundi Onsale Date

SC400/5721/2021

Supervisors' Approval

The undersigned certify that they have read and hereby recommended for acceptance of Murang'a University of Technology a thesis entitled **"An Attention-Based Recurrent Neural Network Model for Improved Aspect-Based Sentiment Analysis of Customer Reviews".**

_____ _____

Dr. Aaron Mogeni Oirere, Date

Department of Computer Science,

Murang'a University of Technology.

_____ _____

Prof. Geoffrey Muchiri Muketha, Date

Department of Computer Science,

Murang'a University of Technology.

DEDICATION

This thesis is dedicated to my beloved parents, Mr. and Mrs. Richard O. Onsale, whose love, sacrifice, and prayers have shaped every step of my journey. To my dear wife, your unwavering support and quiet strength have been a constant source of encouragement. Your presence gave me the courage to persist even during the most challenging moments. I also extend this dedication to my elder brother, Brian Ndege, whose generosity and belief in my goals made this pursuit possible. Each of you has played a unique role in this accomplishment, and I share this success with you.

ACKNOWLEDGEMENT

First and foremost, I express my deepest gratitude to Almighty God, whose grace and presence have been the guiding light throughout this journey. In moments of doubt, He provided reassurance; in times of exhaustion, He renewed my strength. This work would not have been possible without His constant provision, protection, and wisdom. I am truly humbled by the countless ways in which He made a way when none seemed possible. Every milestone reached and every obstacle overcome is a testimony of His faithfulness and unending mercy in my life.

I am profoundly thankful to my supervisors, Dr. Aaron Mogeni Oirere and Prof. Geoffrey Muchiri Muketha, whose mentorship has been instrumental to the success of this thesis. Your thoughtful feedback, timely advice, and academic rigor pushed me to explore deeper dimensions of my research. Dr. Oirere, your patience, and critical insights consistently challenged me to refine my thinking and articulate my ideas with clarity. Prof. Muketha, your encouragement and wealth of experience were invaluable, especially in aligning my research with practical application and academic standards. It has been an honour and a great privilege to work under your supervision and be guided by you both.

I wish to extend heartfelt appreciation to Dr. John Ndia, whose unwavering belief in my potential and consistent encouragement provided a strong pillar of support. I am also deeply grateful to Charles Munyao for his technical insights, Kevin Agina for his thoughtful discussions and moral support, and Jackson Kamiri for always reminding me to keep going, even when things felt overwhelming. My sincere thanks go to my classmates, whose collaboration, friendship, and shared struggles turned this academic endeavour into a memorable and uplifting experience. Each of you played a part in making this journey meaningful and bearable.

Finally, I recognize and appreciate every individual family, friends, faculty, and staff who contributed to my academic growth in any capacity. Your support, no matter how big or small, made a difference. This accomplishment is not mine alone but is shared with all of you who stood by me. Thank you.

ABSTRACT

Sentiment Analysis involves the process of identifying and classifying the underlying opinion or attitude expressed in a piece of text, typically categorizing it as positive, negative, or neutral. Aspect-based sentiment Analysis targets the extraction of sentiments directed toward specific aspects within a review, offering more granular insights than assigning a single sentiment to an entire text. Conventional methods assign a single sentiment label to an entire review but often fail when the text contains mixed opinions about different elements. Deep learning models emerged to address this gap by learning contextual dependencies within sequences, offering improved performance over rule-based approaches. However, despite their strengths, these models often struggle to retain relevant information over long distances, leading to issues such as diluted sentiment signals and reduced sensitivity to aspect-specific content. To overcome these shortcomings, attention mechanisms were integrated to help models focus on crucial sentiment-bearing tokens and better align input elements with aspect terms. However, they still face challenges such as; a limited understanding of complex multi-aspect sentiment where models often fail to accurately interpret multiple opinions spread across different parts of a review, lack of aspect-specific focus as many models treat all words uniformly missing critical phrases linked to specific product features and loss of context in long sequences. A true experimental research design was adopted in this study. The first objective was to compare four deep learning models namely: Recurrent Neural Networks (RNN), Convolutional Neural Network (CNN), Recursive Neural Network (RecNN), and Memory Network (MN). The performance of these models on the SemEval Task 4 dataset was compared without attention mechanisms and with attention mechanisms included. From the performance comparison, RNN integrated with the attention mechanism recorded the highest accuracy score of 84.7%. The second objective was to improve on the above performance index by integrating pretrained multi-head attention and dual attention mechanism into the RNN model, aiming at solving the reported limitations of deep learning models in ABSA. The third objective was to evaluate the performance of the improved model on the SemEval Task 4 and MAMS dataset. On the SemEval Task 4 and MAMS datasets, the improved model achieved accuracy scores of 89.1% and 88.5%, F1-scores of 87.8% and 86.2%, precision scores of 89.6% and 87.3%, and recall scores of 86.7% and 84.2%, respectively. These results indicate consistent improvements across all standard classification metrics compared to the baseline RNN + Hierarchical Attention model. On average across both datasets, the improved model recorded performance gains of 4.1% in accuracy, 3.8% in F1-score, 3.0% in precision, and 4.5% in recall, highlighting the effectiveness of the applied enhancements in improving sentiment classification. These findings demonstrate that the integration of multi-head attention and dual attention mechanisms significantly improved the classification of aspect-specific sentiments in aspect-based customer reviews. The improved model can be deployed in e-commerce platforms, customer feedback systems, and other domains that require detailed, aspect-level sentiment insights to improve products and decision-making.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
LIST OF TABLES ..... **ix**
LIST OF FIGURES ..... **x**
ACRONYMS AND ABBREVIATIONS ..... **xi**
CHAPTER ONE: INTRODUCTION **1**
1.1 Background of the Study 1
1.2 Problem Statement 4
1.3 Research Objectives 4
1.3.1 General Objective 4
1.3.2 Specific Objectives 5
1.4 Research Questions 5
1.5 Significance of the Study 5
1.6 Scope of the Study 6
1.7 Limitations of the Study 6
1.8 Contributions of the Study 6
1.9 Organization of the Thesis 8
CHAPTER TWO: LITERATURE REVIEW **9**
2.1 Introduction 9
2.2 Methods Based on Convolutional Neural Networks (CNNs) 9
2.2.1 Convolutional Neural Networks 9
2.2.2 Deep Learning for Aspect-Based Sentiment Analysis 11
2.2.3 Combination of CNN and GRU for Aspect-Based Sentiment Analysis 12
2.2.4 Ensemble deep learning for aspect-based sentiment analysis 13
2.2.5 Interactive Attention-Based GRU for Aspect Level Sentiment Analysis 14
2.2.6 ABSA Using Deep CNN, LSTM and Fuzzy Logic 15
2.3 Methods Based on Recurrent Neural Networks 16
2.3.1 Recurrent Neural Networks 16
2.3.2 Effect of Epoch Value on RNN-LSTM for Sentiment Analysis 17
2.3.3 Aspect-Based Sentiment Analysis of E-Commerce Reviews 18
2.3.4 Sentiment Analysis Using BDRNN 19
2.4 Methods Based on Recursive Neural Network 20
2.4.1 Recursive Neural Network 20
2.4.2 Combination of RecNN and RNN for ABSA 21

2.5 Advancements to Deep Learning Models 22
2.5.1 Aspect-Level SA Using CNN, Bi-LSTM, and Attention Mechanism 22
2.5.2 Learning to Attend via Word-Aspect Associative Fusion for ABSA 23
2.5.3 Embedding Commonsense Knowledge into Attentive LSTM 24
2.5.4 Aspect-Based Sentiment Analysis Using Attention-based GRU 25
2.5.5 Positionless Aspect-Based Sentiment Analysis Using Attention 26
2.5.6 Hyperparameter Optimization in Deep Learning Models 26
2.6 Evaluation Metrics 28
2.6.1 Accuracy 28
2.6.2 Precision 29
2.6.3 Recall 30
2.6.4 F1-Score 30
2.6.5 Macro-Averaged Precision, Recall, and F1-Score 31
2.6.6 Micro-Averaged Precision, Recall, and F1-Score 32
2.6.7 Weighted-Averaged Precision, Recall, and F1-Score 33
2.6.8 Confusion Matrix 34
2.6.9 Aspect Detection Accuracy 35
2.7 Summary 36
CHAPTER THREE: METHODOLOGY 38
3.1 Introduction 38
3.2 Research Design 38
3.3 Research Process 39
3.4 Data Acquisition 40
3.4.1 Dataset Description 40
3.4.2 Data Pre-processing 42
3.5 A Comparative Study of Existing Deep Learning Models 43
3.6 Developing the Improved Model 45
3.6.1 Multi - Head Hierarchical Attention Mechanism 45
3.6.2 Dual Attention Mechanism for Aspect-Specific Focus 51
3.6.3 Hyperparameter Optimization 55
3.6.4 Model Training and Validation 57
3.7 Model Evaluation 60
3.8 Experimental Materials 62
3.9 Ethical Considerations 63
3.10 Summary 63
CHAPTER FOUR: RESULTS AND DISCUSSION 65
4.1 Introduction 65
4.2 A Comparative Study of Existing Deep Learning Models 65
4.2.1 Baseline models without attention mechanisms 65

4.2.2 Baseline Models with Attention Mechanisms 67
4.2.3 Training and Validation Loss of the Models 68
4.3 An Improved Deep Learning-based Model 73
4.3.1 Integrated Multi-head Attention to Hierarchical Attention Mechanism 74
4.3.2 Dual Attention Mechanism for Aspect-Specific Focus 75
4.3.3 Hyperparameter Optimization 75
4.3.4 Improved Model Architecture 76
4.4 Evaluation of Baseline and the Improved Model 79
4.4.1 Model Performance on the four Standard Metrics 79
4.4.2 Training and Validation Loss of the Model 82
4.4.3 Comparing the Improved Model with the Existing Models 83
4.5 Discussion 84
4.5.1 Implications of the Comparative Study of Deep Learning Models 85
4.5.2 Implications of the Improved Model 86
4.5.3 Implications of the Evaluation Results of the Improved Model 88
4.6 Summary 90
CHAPTER FIVE: CONCLUSION, RECOMMENDATIONS AND FUTURE WORK 91
5.1 Conclusion 91
5.2 Recommendations 93
5.2.1 Policy Support in Customer Feedback Systems Deployment 93
5.2.2 Institutional Adoption in Customer Experience Monitoring 93
5.2.3 Cross-Sector Training and Capacity Building 93
5.3 Future Work 94
5.3.1 Multimodal Integration 94
5.3.2 Cross-Lingual Sentiment Analysis 94
5.3.3 Real-Time Sentiment Analysis Deployment 95
REFERENCES 96
APPENDICES 107

LIST OF TABLES

Table 3.1: SemEval Task 4 Dataset Distribution 41
Table 3.2: Multi-Aspect Multi-Sentiment Dataset 42
Table 3.3: Experimental Materials 62
Table 4.1: Deep Learning Models without Attention Mechanism 66
Table 4.2: Performance of Baseline Deep Learning Models with Attention 68
Table 4.3: Optimized Hyperparameters for the Improved RNN 76
Table 4.4: Evaluation Results Using Standard Classification Metrics 80
Table 4.5: Improved Model with Existing Models on the SemEval Task 4 84
Table 4.6: Improved Model with Existing Models on the MAMS Dataset 85

LIST OF FIGURES

Figure 1.1: An Example of Aspect Level Classification 3
Figure 2.1: CNN-based Model for Sentence Classification 10
Figure 3.1: The Research Process 40
Figure 3.2: Baseline RNN Architecture 46
Figure 3.3: Baseline RNN with Hierarchical Attention 47
Figure 3.4: Integrated Multi-Head Attention Mechanism 49
Figure 3.5: Dual Attention Mechanism for Aspect Specific Focus. 53
Figure 3.6: Hyperparameter Optimization 57
Figure 4.1: Training and Validation Loss - RNN 70
Figure 4.2: Training and Validation Loss - CNN 71
Figure 4.3: Training and Validation Loss - RecNN 72
Figure 4.4: Training and Validation Loss - Memory Networks 73
Figure 4.5: Architecture of the Improved RNN Model 78
Figure 4.6: The SemEval Confusion Matrix 81
Figure 4.7: MAMS Confusion Matrix 82
Figure 4.8: Training and Validation Loss - Improved RNN+Hierarchical 83

ACRONYMS AND ABBREVIATIONS

ACC	Aspect Category Classification
ABSA	Aspect-Based Sentiment Analysis
ACSA	Aspect Category Sentiment Analysis
ASC	Aspect Sentiment Classification
ATSA	Aspect Term Sentiment Analysis
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DNNs	Deep Neural Networks
GRU	Gated Recurrent Unit
IAC-GRU	Interactive Attention Convolutional Gated Recurrent Unit
LSTM	Long Short-Term Memory
MAMS	Multi-Aspect Multi-Sentiment
MN	Memory Network
NLP	Natural Language Processing
RecNN	Recursive Neural Network
RNN	Recurrent Neural Network
BDRNN	Bidirectional Deep Recurrent Neural Networks
AC-CNN-SA	Attention-Controlled Convolutional Neural Network
NACOSTI	National Commission for Science, Technology and Innovation
AF-LSTM	Aspect Fusion LSTM
MAE	Mean Absolute Error
RMAE	Root Mean Square Error

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

User-generated content on various online platforms is increasing rapidly, offering valuable insights that can significantly support decision-making processes. However, the enormous amount of data available makes it very difficult to achieve accurate information extraction [1], [2]. Many studies in Natural Language Processing have focused on sentiment analysis as a means to sift through the mountain of user reviews. Views, feelings, evaluations, attitudes, and emotions are just a few of the numerous textual aspects that are investigated in the area of opinion mining, also known as subjectivity analysis [3]. Data mining and information retrieval are the roots of the term "opinion mining," which is often used to describe sentiment analysis [4]. The primary goal is to ascertain the general public's stance on certain subjects. Recognizing and categorizing ideas expressed in text as positive, negative, or neutral is an important function of sentiment analysis.

Extensive research has emphasized the significant level of confidence that customers have in reviews and remarks from unknown persons when considering purchasing products or services. There have been many extensive statistical surveys and studies conducted in the field of online customer feedback [5], [6]. The research provided in [7] reveals that prospective consumers have a significant inclination to invest an additional 31% of their expenditure towards items or services that have outstanding ratings. Another research that highlights the importance of customer reviews is the one in [8]. It shows that when given an option between two items, consumers aren't inclined to choose the one with fewer reviews. As seen in [9], an outstanding 98% of consumers are reluctant to acquire items that have little or no evaluations. Furthermore, as explained in [10], when bad reviews appear, approximately 80% of consumers rethink purchasing items that have been suggested to them by friends or family.

Three main types of sentiment analysis are aspect-based, sentence-based, and document-based [11], [12], [13]. Assuming all documents deal with the same subject, the simplest method, Document-Based Sentiment Analysis, finds out whether the whole document expresses a positive or negative review. Nevertheless, a more

sophisticated examination is required when papers assess multiple entities since this presumption falls short.

In Sentence-Based Sentiment Analysis, the goal is to discern whether individual sentences express positive, negative, or neutral opinions about an entity, introducing the concept of neutrality absent in document-based analysis [14],[15]. Like its predecessor, it also presumes a single entity per sentence. Nonetheless, these first two levels excel when documents or sentences revolve around a single entity. Conversely, Aspect-Based Sentiment Analysis delves into the most granular level of analysis, typically found in product reviews on platforms like Amazon, Yelp, or eBay [16], [17]. It pertains to products with multifaceted attributes, like cars, cameras, and mobile phones, where people express distinct opinions on various aspects of the entity.

To conduct sentiment analysis, one must first identify the quadruple (S, G, H, and T) where S stands for sentiment, G for the entity that the sentiment is directed towards, H for the person or thing expressing the sentiment, and T for the time stamp of the sentiment expression [18], [19]. Identifying the sentiment-entity pair (S, G) is the primary emphasis of most techniques. In this setting, the nature of the entity changes depending on the area of application; for example, in product evaluations, the product is usually the entity, and its features, such as quality and price, are aspects.

The goal of Aspect-Based Sentiment Analysis (ABSA) is to determine the general feeling about something and then break it down into the feelings about its individual parts. It becomes really important to have this level of detail when a popular product has a lot of reviews and it's hard for people to read them all before buying [20], [21]. At the same time, after receiving client input, ABSA gives manufacturers extensive insights to improve certain parts of their goods. "Powerful Engine for a Subaru but High Maintenance Costs" is one such example as shown in Figure 1.1. The consumer expresses satisfaction with the product's "strength" but strong disapproval of its "maintenance cost" in this response.

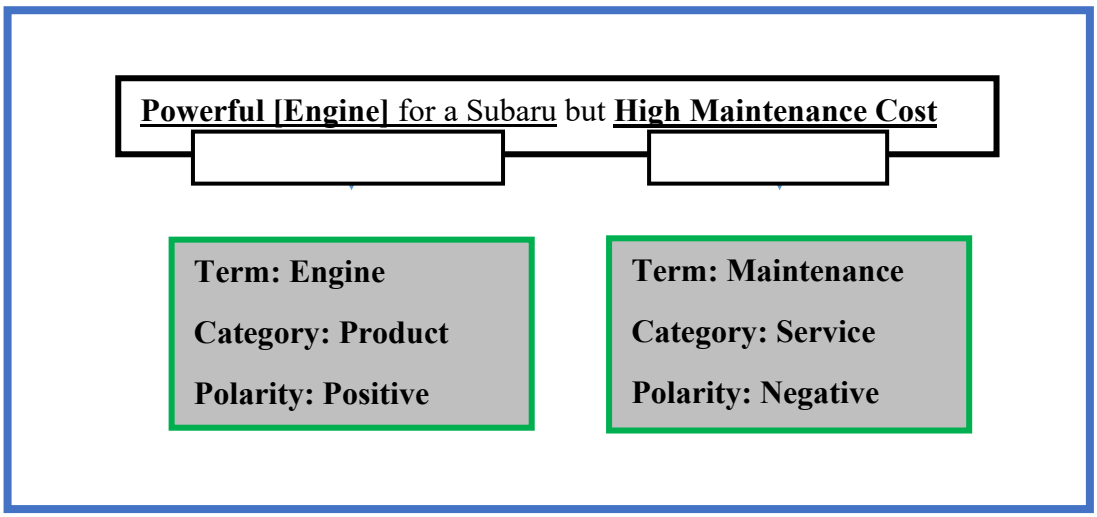


Figure 1.1: An Example of Aspect Level Classification

The journey of sentiment analysis began with rule-based systems that used hand-crafted language rules and dictionaries to assign opinions to text [22]. These systems worked well for basic sentences but failed when the text became more complex. They couldn't handle informal language, mixed opinions, or the deeper meaning behind words. As data grew, researchers shifted to machine learning methods [23]. Tools like Support Vector Machine (SVM) and Naive Bayes improved classification by learning patterns from labelled data [24], [25]. However, these models relied heavily on manual features and lacked the ability to understand context or detect multiple opinions in the same review.

Later, deep learning brought a major shift by removing the need for hand-crafted features. Models like Recurrent Neural Network (RNNs), Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTMs) helped capture word sequences and hidden meanings better than earlier methods [26]. They improved performance but had trouble with longer texts, often forgetting important details from earlier parts of a sentence [27]. To solve this, attention mechanisms were introduced to help models focus on key words or phrases [28], [29]. These mechanisms improved how models captured aspect-level opinions in reviews. Still, some challenges remain. Not all attention types are equally strong, and many models struggle to correctly identify opinions linked to different product features [30]. This leaves room for better solutions that combine deep learning with more refined attention techniques.

1.2 Problem Statement

Capturing sentiment in written text is essential for drawing meaningful insights from customer reviews, feedback, and social media content. However, accurately detecting sentiment becomes a challenge, especially when multiple or conflicting opinions appear within the same text, leading to misclassification, which limits the reliability of existing sentiment analysis approaches [31].

Over time, sentiment analysis has evolved. Early approaches relied on rule-based and lexicon-driven methods, which were easy to implement but struggled with context and ambiguity. Conventional machine learning techniques, such as SVMs and Naive Bayes, improved performance by learning from data, yet they required extensive feature engineering and still lacked deep contextual understanding. Deep learning models were introduced to address conventional techniques challenges by automatically learning features and capturing sequential patterns [34]. However, these models often failed to retain key information across long sequences and treated all parts of the text with equal weights, reducing their effectiveness in more complex sentiment tasks [36].

To address deep learning limitations, attention mechanisms were integrated into deep learning models, enabling them to focus on the most relevant parts of a sentence or review. Even though this integration enhanced the performance of deep learning models, they still struggle with aspect-based sentiment analysis, particularly when sentiments are spread across different parts of a review or when multiple aspects are included [45]. This struggle is manifested in the lack of aspect-specific focus, given that these models treat all words uniformly, consequently missing critical phrases linked to specific product features [42]. This results in the loss of context in long sequences while such models struggle to retain relevant information when sentiments span lengthy text [44]. These limitations show that the underlying problem is not fully solved yet and that it still persists.

1.3 Research Objectives

1.3.1 General Objective

The general objective of this study was to develop an attention-based recurrent neural network model for improved aspect-based sentiment analysis of customer reviews.

1.3.2 Specific Objectives

i. To determine architectural features that can improve the performance of a suitable deep learning for aspect-based sentiment analysis.
ii. To develop an improved deep learning model incorporating attention mechanisms tailored for aspect-based sentiment analysis of customer reviews.
iii. To evaluate the performance of the improved model on the aspect-based sentiment analysis benchmark datasets.

1.4 Research Questions

i. What is the most suitable baseline model and which architectural features can enhance deep learning performance in aspect-based sentiment analysis?
ii. How can integrating various attention mechanisms improve the performance of deep learning models for aspect-based sentiment analysis of customer reviews?
iii. How can the performance of the improved model be evaluated on the aspect-based sentiment analysis benchmark datasets?

1.5 Significance of the Study

This study contributed to the ongoing development of advanced sentiment analysis techniques by focusing on aspect-based sentiment analysis using deep learning. Its core value lies in improving how customer opinions are understood and categorized at a more detailed level, particularly when reviews cover multiple product features. Businesses can benefit from this work by gaining more accurate insights into specific aspects of their products or services, which supports informed decision-making, better marketing strategies, and targeted product enhancements.

The study also delivered an improved model that provides value to researchers in natural language processing by addressing key limitations in existing models such as poor handling of long text, lack of aspect focus, and loss of contextual information while demonstrating how enhancements like hierarchical and multi-head attention can improve performance. The insights from this research can be extended to other domains where a precise understanding of user feedback is essential, such as healthcare, education, or finance, making it a significant contribution to both industry and academic fields.

1.6 Scope of the Study

This study falls under the field of Natural Language Processing, a key area in Artificial Intelligence. It focused on aspect-based sentiment analysis, using the SemEval task 4 and Multi-Aspect Multi-Sentiment datasets particularly the laptop reviews. From the wide range of ABSA applications across different languages and domains, this research specifically targeted customer opinions on product features, specifications, and pricing. The study involved identifying suitable deep learning models through a comparative study and improving them by incorporating multi-head attention, dual attention mechanisms and the identified features. The improved model demonstrated consistent gains in classification metrics - accuracy, precision, recall, and F1-score on the benchmark datasets making it competitive with state-of-the-art systems. These improvements have academic value in refining attention-based model design and practical significance in supporting product feedback analysis, customer experience optimization, and sentiment monitoring in e-commerce platforms. The model's architecture can be adapted for real-time or production-level deployment in industry environments.

1.7 Limitations of the Study

This study relied exclusively on secondary data sources, specifically the SemEval Task 4 and Multi-Aspect Multi-Sentiment datasets for training and evaluating the model. These datasets are widely recognized in aspect-based sentiment analysis research and provided a solid foundation for testing the proposed architecture. However, the study did not involve collecting new data, which means the analysis was based on predefined review samples. As a result, the study focused on improving model performance within the scope of those datasets and did not explore additional product categories or language variations that could be considered in future work.

1.8 Contributions of the Study

1. **Theoretical Contributions to the Field of Knowledge**
    - **Advancement of ABSA Models:** This study enhances the theoretical framework of aspect-based sentiment analysis (ABSA) by integrating advanced attention mechanisms into deep learning models [include the specific mechanisms from this study]. It provides empirical evidence that attention-based architectures can better capture

sentiment nuances tied to specific aspects, contributing to the understanding of model interpretability in NLP.

- **Feature Importance Insights:** Through identifying and analysing key features that impact model performance, the research offers valuable insights into the role of linguistic and contextual elements in sentiment analysis. This contributes to the broader theoretical discourse on feature engineering in deep learning.
- **Interdisciplinary Integration:** The study bridges theories from computational linguistics, cognitive science, and marketing, fostering interdisciplinary dialogue. This integration enriches the theoretical landscape of sentiment analysis by emphasizing the importance of context and human cognition in language processing.

2. **Contributions to Practice**
    - **Enhanced Model Performance:** The development of a robust, attention-based model for ABSA directly benefits practitioners in fields such as marketing, customer service, and product management. Companies can leverage this model to gain deeper insights into customer sentiments, leading to more informed decision-making.
    - **Improved Customer Insights:** Through accurately identifying sentiments related to specific aspects of products or services, businesses can tailor their strategies to address customer concerns, improve product offerings, and enhance customer satisfaction.
    - **Practical Model for Implementation:** The study provides a clear model for implementing advanced sentiment analysis models in real-world applications, guiding practitioners on feature selection, model training, and evaluation processes.
3. **Contributions to Policy**
    - **Informed Decision-Making for Stakeholders:** The findings can inform policymakers and stakeholders about the potential of advanced sentiment analysis tools in shaping customer engagement strategies. Through understanding consumer sentiments, policies can be adapted to better reflect public opinion and enhance service delivery.

- **Guidelines for Ethical AI Use:** The study highlights the importance of ethical considerations in sentiment analysis, particularly regarding data privacy and user consent. This contributes to the development of guidelines for ethical AI practices in data collection and analysis, influencing policy decisions in organizations.
- **Regulatory Frameworks:** As organizations increasingly rely on sentiment analysis for decision-making, the study's insights can contribute to the formulation of regulatory frameworks that ensure transparency and accountability in the use of AI technologies.

1.9 Organization of the Thesis

This thesis is organized into five chapters explained below:

Chapter one provides the background to the study, defines the research problem, outlines the research objectives, highlights the significance of the study, and states the scope, limitations, and contributions of the thesis.

Chapter two reviews existing literature on sentiment analysis, aspect-based sentiment analysis, deep learning models, attention mechanisms, and identifies the gaps that this study seeks to address.

Chapter three describes the research methodology, including the research design, research process, dataset acquisition, data preprocessing steps, model development process, experimental materials, and ethical considerations.

Chapter four presents the results obtained from the baseline models and the improved model, and discusses the findings based on evaluation metrics such as accuracy, precision, recall, and F1-score.

The fifth chapter provides the conclusions drawn from the study, gives recommendations based on the findings, and suggests future research directions for further improvements in aspect-based sentiment analysis.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents a thorough review of scholarly works relevant to this study. It begins by outlining the review strategy and structure, starting with convolutional neural network (CNN) approaches, followed by recurrent neural networks (RNNs) and their variants such as LSTMs and GRUs, then recursive neural networks (RecNNs), memory networks, and finally attention-based innovations [46], [47], [48], [50]. The literature is organized chronologically and thematically to highlight the evolution of deep learning techniques in sentiment analysis. Particular focus is placed on models commonly applied in Aspect-Based Sentiment Analysis (ABSA), examining how these have been adapted and improved to better handle aspect detection and sentiment classification. Key enhancements to model architectures are discussed, along with evaluation metrics widely used to assess ABSA performance. The chapter concludes by identifying critical research gaps that the present study seeks to address.

2.2 Methods Based on Convolutional Neural Networks (CNNs)

Convolutional Neural Networks have been widely adopted in ABSA due to their efficiency in capturing local patterns within text as discussed in the following subsections;

2.2.1 Convolutional Neural Networks

Convolutional Neural Networks first gained prominence in the field of computer vision, where they transformed tasks like image recognition and classification [51], [52]. However, starting around 2011, researchers in Natural Language Processing began successfully applying CNNs to a range of language tasks, including sentiment analysis, machine translation, and question answering [53], [54]. Unlike image-based applications, CNN models in NLP work directly with text. To process textual data, each input sentence is converted into a matrix format, where each row corresponds to a word and each word is represented by a vector of fixed dimensions [55]. For instance, consider the sentence: “Good for COVID-19 but mask material is uncomfortable.” This sentence contains eight words, and if each word is mapped as a five-dimensional vector, the input becomes an 8x5 matrix [49]. As illustrated in Figure

2.1, a typical CNN model for NLP includes layers such as the input layer, convolutional filters, max-over-time pooling, and a final SoftMax layer for classification.

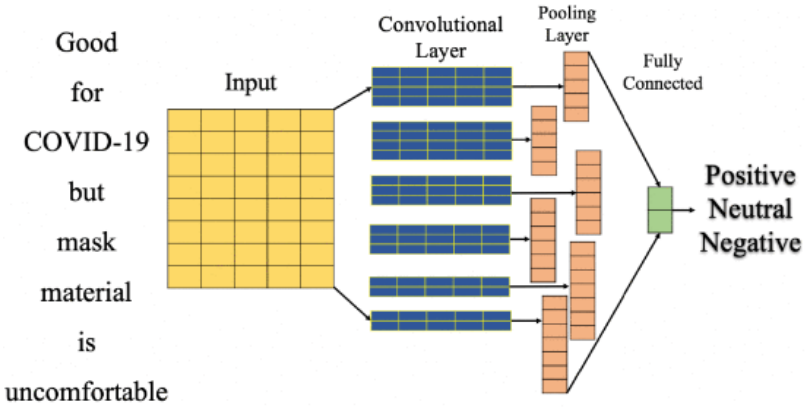


Figure 2.1: CNN-based Model for Sentence Classification

The input layer transforms the sentence into a structured matrix, where each row represents a word and each word is encoded as a *k*-dimensional vector. For a sentence containing *n* words, the input matrix has dimensions $n \times k$. This grid-like structure allows the model to preserve the word order and maintain local context for each word across the sentence.

The convolutional layer applies multiple filters over the input matrix using a sliding window technique. Each filter has a window size of $h \times k$, where h defines how many consecutive words are covered (height), and *k* matches the dimensionality of each word vector (width). The convolution operation produces a feature c_i for every window, calculated as:

$$c_i = s(w \cdot x_{i:i+h-1} + b)$$

Here,

w is the filter's weight matrix

$x_{i:i+h-1}$ is the concatenated vector of words from position *i* to *i+h-1*

b is the bias term

s is a nonlinear activation function, such as tanh or ReLU

Sliding the filter across all positions in the sentence, we obtain a feature map:

$$c = [c_1, c_2, \dots, c_{n-h+1}]$$

This map highlights key local patterns captured from adjacent words. The pooling layer condenses the feature map into a single representative value per feature, typically using max-pooling. This operation selects the maximum value from each feature map, ensuring the most salient feature is preserved:

$$\hat{c} = \max\{ c_1, c_2, \dots, c_{n-h+1}\}$$

This reduction step helps control model complexity, improve generalization, and retain the strongest signals detected by the filters. Finally, the pooled features are passed to a fully connected layer, which feeds into a SoftMax classifier. The SoftMax function transforms the raw scores into normalized probabilities across the target classes (e.g., Positive, Neutral, Negative):

$$P(y = j \mid \hat{c}) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

Where;

z_j is the score for class j

K is the total number of sentiment classes.

This layer outputs the model's predicted sentiment category based on the highest probability.

2.2.2 Deep Learning for Aspect-Based Sentiment Analysis

Bo Wang and Min Liu [56] proposed a deep learning framework for aspect-based sentiment analysis that integrates two specialized models: one for aspect identification and another for sentiment prediction. For the aspect model, they employed a two-layer neural network using pretrained 300-dimensional Google News word embeddings to represent each sentence, applying a softmax classifier to predict the probability distribution over 19 aspect categories. Importantly, they reduced model complexity by merging infrequent aspect categories into an "OTHER" class and introducing a "NONE" label for sentences without any relevant aspect, allowing the system to handle domain-specific variations in the laptop review dataset more effectively.

The researchers moved beyond simple bag-of-words approaches by adopting a convolutional neural network designed to work at the sentence level for sentiment prediction. To connect aspects with their corresponding sentiments, they introduced an innovative rescaling method where each word vector was adjusted according to its relation to the targeted aspect, using parse-tree-based probabilistic propagation. Their experiments on the SemEval-2015 dataset showed that the aspect model achieved an F1 score of 0.513, outperforming the competition winners, while the sentiment model achieved 78.3% accuracy, tying for second place.

Despite these successes, the study highlighted a key challenge tied to the lack of aspect-specific focus: even with their sophisticated rescaling mechanism, the system sometimes struggled to isolate and interpret sentiment cues accurately for each individual aspect when reviews contained mixed or overlapping opinions. This emphasized the need for further research into architectures capable of providing stronger aspect-level discrimination in multi-aspect sentiment settings.

2.2.3 Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis

Zhao et al. [57] explored the combination of Convolutional Neural Networks and Gated Recurrent Units to improve aspect-based sentiment analysis. Their hybrid model was designed to harness CNNs' strengths in local pattern detection and GRUs' ability to preserve sequential dependencies over time. The researchers initialized the system using pretrained word2vec embeddings and carefully tuned hyperparameters, achieving promising results across datasets from the automobile and hotel domains. Their approach demonstrated notable improvements in emotion classification and aspect extraction by leveraging CNNs to capture n-gram-level features and GRUs to carry forward contextual signals across sentence boundaries.

An important methodological insight from their work was the sensitivity to word embedding dimension. Zhao et al. reported that embedding sizes that were too small led to underfitting, while overly large embeddings caused overfitting and increased computational demands without improving accuracy. This emphasized the need for careful balancing between representational richness and model complexity. The study also raised the prospect of developing models that can process multiple aspects

simultaneously, suggesting that future architectures might benefit from integrating attention mechanisms or transformer-based structures.

Although the CNN-GRU model marked progress in achieving 83.6% on the Area Under the Curve, it still faced clear limitations, particularly in terms of limited understanding of complex context and multi-aspect sentiment. Despite combining CNNs and GRUs, the model struggled to capture distant sentiment signals spread across lengthy or multifaceted reviews. This limitation meant that when faced with texts containing multiple intertwined opinions, the model risked missing important relationships between aspects, leading to incomplete or imprecise sentiment classification.

2.2.4 Ensemble deep learning for aspect-based sentiment analysis

Mohammadi and Shaverizade [58] introduced an innovative approach to aspect-based sentiment analysis using deep ensemble learning. Their study began by independently constructing four distinct deep learning models: Gated Recurrent Units, Long Short-Term Memory networks, Bidirectional LSTM, and Convolutional Neural Networks. The novelty of their work lay in employing a stacking ensemble method, where the outputs of these individual models were merged, and a logistic regression layer served as the meta-learner to produce the final predictions. This approach was rigorously tested on datasets from the laptop and restaurant domains. To ensure robust training, the authors applied early stopping to all models, preventing overfitting during the 30 training epochs executed with a batch size of 32. Performance comparisons were drawn against baseline classifiers such as standalone CNN, BiLSTM, LSTM, GRU, and other simpler models.

While the ensemble approach delivered improved precision reaching about 69.3% on the Restaurant Dataset and 67.5% on the Laptop Dataset the results revealed that the GRU component consistently underperformed compared to the others. Additionally, the memory-based models in the ensemble encountered notable difficulties when tasked with handling reviews where the sequence of sentences changed, which impaired their ability to trace sentiment relationships between different aspects. This limitation aligns directly with the loss of context in long sequences challenge, as the models struggled to maintain accurate sentiment interpretations when key contextual cues spanned across reordered or extended sentence structures. The findings suggested

that, despite the ensemble's promising design, further enhancements were necessary to strengthen its capacity to retain and interpret sentiment cues over complex, multi-sentence inputs.

2.2.5 Interactive Attention-Based Convolutional GRU for Aspect Level Sentiment Analysis

Chen et al. [59] introduced the Interactive Attention-Based Convolutional Bidirectional Gated Recurrent Unit model, offering a sophisticated blend of recurrent and convolutional architectures for aspect-level sentiment analysis. The model integrated RNN-based sequence modelling with CNN-driven feature extraction while embedding a bidirectional GRU within the convolutional framework to enhance the depth of contextual understanding. A key innovation of the IAC-GRU approach was its use of positional and Part-of-Speech information within the embedding layer, allowing the system to treat each target (or aspect) independently and with greater sensitivity to syntactic and positional cues. These enhancements resulted in improved sentiment classification performance of 76.2%, particularly when compared to earlier standalone models, positioning IAC-GRU as a promising step forward in the field.

Despite its achievements, the IAC-GRU model exhibited a notable limitation tied to the lack of aspect-specific focus. While the model succeeded in improving general sentiment prediction, it did not fully resolve the challenge of distinguishing which sentiment phrases corresponded to which specific product aspects. This gap became apparent in scenarios involving multi-aspect reviews, where overlapping or interacting sentiments required fine-grained, aspect-sensitive interpretation. Although the model enhanced overall accuracy, it fell short of providing the targeted, aspect-level attention necessary for precise opinion extraction, especially when multiple product features were discussed within the same review. Addressing this shortcoming calls for more refined architectures capable of assigning differentiated importance to words or phrases depending on the targeted aspect an objective that the present research tackled by incorporating advanced hierarchical and multi-head attention mechanisms.

2.2.6 Aspect-Based Sentiment Analysis of Mobile Phone Reviews Using Deep CNN, LSTM, and Fuzzy Logic

Murugaiyan and Reddy [60] presented a hybrid framework combining Deep Convolutional Neural Networks, Bidirectional Long Short-Term Memory networks (BiLSTM), and fuzzy IF-THEN rule-based logic to perform aspect-based sentiment analysis of mobile phone reviews. Their system processed both audio and text data by first extracting Mel Frequency Cepstral Coefficients from customer speech inputs, followed by DCNN-based emotion recognition, and finally using BiLSTM to identify the relevant product aspects, such as battery life or display quality. The fuzzy logic module then classified the sentiment linked to each aspect as positive, negative, or neutral. This combination allowed the system to harness the strengths of deep learning models for feature extraction and sequence modelling while leveraging the interpretability of fuzzy logic for final decision-making. The authors tested their system on benchmark datasets like RAVDESS, SAVEE, EMO-DB, and IEMOCAP and reported impressive classification accuracy that often surpassed the performance of existing models.

Despite these advances, the study faced notable constraints related to its reliance on predefined fuzzy rules, which limits the model's flexibility when applied to reviews containing multiple, layered opinions. While the fuzzy IF–THEN system provided clarity and interpretability, it struggled to handle cases where sentiments about different aspects were spread across several sentences or expressed in more complex linguistic patterns. Additionally, the architecture's dependence on BiLSTM for aspect detection did not fully resolve the challenge of recognizing relationships between distant words or sentences, particularly when opinions shifted between multiple product features within the same review. This makes the system vulnerable to incomplete or inaccurate sentiment classification, especially in settings where user feedback is rich, detailed, and interwoven with mixed evaluations of various aspects.

The key shortcoming identified here aligns with the challenge of limited understanding of complex context and multi-aspect sentiment. While the hybrid approach improved baseline performance, it did not offer a robust solution for capturing the intricate links between multiple sentiments and aspects in long or detailed reviews. This highlights a critical gap in the architecture, as it lacked

mechanisms like hierarchical or multi-head attention that could enable the model to dynamically weigh different parts of the input when analysing sentiment relationships. Addressing this weakness is essential for developing models that not only improve classification metrics but also provide a more accurate and nuanced understanding of customer feedback.

2.3 Methods Based on Recurrent Neural Networks

Recurrent Neural Networks have gained popularity in ABSA due to their ability to model the sequential nature of language, where sentiment often unfolds over time. As discussed in the following subsections, RNN variant such as LSTM was developed to address the limitations of basic RNN by preserving contextual information.

2.3.1 Recurrent Neural Networks

RNNs have played a foundational role in the development of aspect-based sentiment analysis by offering a way to model sequential data with dependencies across time or word positions [61]. Unlike feedforward networks, RNNs are designed to handle inputs where the current output depends not only on the present input but also on previous computations [62]. A sample RNN structure is shown in Figure 2.2.

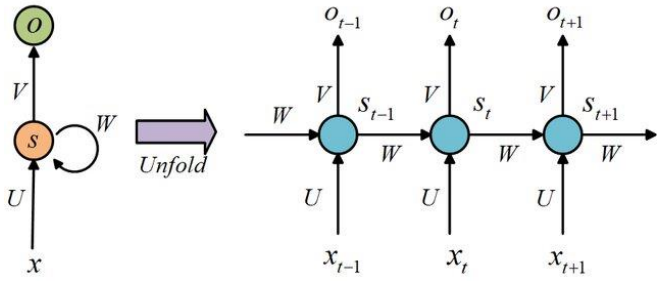


Figure 2.2: Basic RNN Structure

As illustrated in Figure 2.2, the RNN architecture consists of a looping structure where the hidden state at time step t, denoted as s_t, is computed based on the current input x_t and the hidden state from the previous step s_{t-1}. Mathematically, this can be expressed as:

$$s_t = f(Ux_t + Ws_{t-1})$$

Where;

U and W are weight matrices

f is a nonlinear activation function such as tanh or ReLU.

This recursive formulation allows the network to "remember" information from earlier parts of a sequence, making it particularly useful for tasks like sentiment analysis where the meaning of a phrase often depends on its context within the sentence. Despite their usefulness, standard RNNs face challenges when applied to longer text sequences due to the problem of vanishing or exploding gradients, which weakens their ability to capture long-range dependencies. To produce meaningful outputs, the RNN computes an output o_t at each time step, typically as:

$$o_t = g(Vs_t)$$

Where;

V is the output weight matrix

g is an activation function suited for the task, often softmax for classification.

In ABSA, this enables the model to make predictions about sentiment at each aspect-relevant position in the text. However, the limitations of standard RNNs have led researchers to explore advanced variants like Long Short-Term Memory Networks and Gated Recurrent Units, which introduce gating mechanisms to better manage the flow of information over extended sequences setting the stage for further improvements.

2.3.2 Effect of Epoch Value on RNN-LSTM Performance for Sentiment Analysis

Maswan Pratama Putra and Yuliant Sibaroni [63] investigated how varying the epoch value affects the performance of an RNN-LSTM model in classifying sentiment from Lazada e-commerce reviews. They gathered over 5,900 Indonesian-language user reviews using automated crawling tools and manually labelled the dataset into positive and negative categories. Before feeding the data into the model, they applied two feature extraction techniques: Term Frequency–Inverse Document Frequency and Word2Vec embeddings, which transformed raw textual input into numerical vectors suitable for deep learning. The RNN-LSTM model, known for its ability to capture sequential dependencies through specialized memory gates, was then trained under

different configurations to determine which settings maximized performance. A sample TFIDF formula used is shown below.

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

The formula computes the importance of term *t* in document *d*,

Where;

N is the total number of documents

DF(t) is the number of documents containing *t*.

Their experiments revealed that the combination of five training epochs, a dropout rate of 0.8, and a batch size of 32 produced the highest accuracy, reaching 86.18%. Importantly, this configuration balanced the precision, recall, and F1 scores, demonstrating the model's ability to consistently classify e-commerce reviews across both sentiment classes. The authors emphasized that the LSTM's design, which allows it to preserve long-term dependencies better than a vanilla RNN, was central to achieving these strong results. However, they also acknowledged that further hyperparameter tuning and experimentation were essential to push the model toward even higher performance, especially on larger or more diverse datasets.

Despite these promising outcomes, the study highlighted a persistent challenge connected to loss of context in long sequences. Specifically, the authors observed performance drops when the reviews contained words or expressions not well represented by the pretrained Word2Vec embeddings. This mismatch affected the LSTM's ability to maintain accurate sentiment tracking over extended texts, particularly when critical context was dispersed across multiple sentences or paragraphs. These findings underscored that, beyond parameter tuning, effective ABSA requires models that can dynamically adjust to shifting vocabularies and sustain contextual awareness over long passages a limitation this research sought to address through advanced attention-based improvements.

2.3.3 Aspect-Based Sentiment Analysis of Customer Reviews in E-Commerce

In their recent study, Emmanuel Chris et al. [64] conducted an in-depth investigation into how aspect-based sentiment analysis can be effectively applied to customer

reviews within e-commerce platforms. The authors emphasized the importance of moving beyond simple positive or negative classifications by implementing fine-grained analyses that separate sentiments for specific product aspects, such as camera quality, battery life, or delivery experience. To achieve this, the study leveraged recurrent neural networks, particularly long short-term memory architectures, to process the sequential nature of text data, making it possible to detect contextual relationships across words and identify sentiment polarity for each detected aspect.

The authors designed their model pipeline with three major components: aspect extraction, sentiment classification, and opinion polarity detection. They used dependency parsing and named entity recognition methods for precise extraction of opinion targets and applied LSTM-based models for sentiment classification. Their results demonstrated that RNN-based models offered superior contextual understanding compared to traditional machine learning classifiers. However, the study also acknowledged the challenge of implicitly expressed aspects, where features like “battery life” are hinted at without being explicitly mentioned, indicating that further refinements are required to improve implicit aspect detection accuracy.

Notably, the paper raised concerns about the model’s limited ability to capture long-range dependencies when multiple aspects were dispersed across lengthy reviews. This aligns with one of the challenges identified in this research - the problem of losing important context in long sequences, especially when RNNs encounter difficulty in maintaining the necessary memory of earlier tokens. Addressing this limitation is critical for ensuring accurate sentiment analysis across multiple aspects, particularly when the relationships between sentences span several product features.

2.3.4 Sentiment Analysis Using Bidirectional Deep Recurrent Neural Networks (SA-BDRNN)

Dr. Muthusankar et al. [65] proposed a sentiment analysis framework called SA-BDRNN (Sentiment Analysis–Bidirectional Deep Recurrent Neural Networks) designed to handle large-scale social media data and assess public opinion. This system was structured into four main stages: constructing sentiment word repositories, classifying emotional behaviour using BDRNN, balancing extended words, and predicting sentiment polarity. Their work uniquely incorporated a reduced, carefully selected set of hashtags, refined through tokenization, stemming, and filtering, to build

an annotated repository capable of capturing refined sentiments linked to topics like the Jallikattu protests. Compared to earlier frameworks like CNN-DLSA, AC-CNN-SA, and TSA-DNNM, the SA-BDRNN scheme showed superior accuracy, precision, recall, F-measure, and lower loss rates, thanks to its tailored preprocessing and robust BDRNN architecture.

The proposed model utilized both forward and backward hidden layers to capture past and future contexts simultaneously, making it highly effective for sequential text tasks. Key system components included balancing of class weights, calculating the degree of polarity and impact, and aggregating these to produce a reliable final sentiment rating. The study's experiments, which processed over 3.2 million tweets, demonstrated consistent performance improvements across metrics such as MAE, RMSE, Kappa statistics, and true positive rate. These gains highlighted the importance of integrating detailed linguistic preprocessing and bidirectional architectures in real-time social media sentiment analysis pipelines.

Despite these achievements, the study implicitly pointed to the challenge of lack of aspect-specific focus. The SA-BDRNN system emphasized overall sentiment polarity across posts but did not explicitly differentiate between sentiments related to individual aspects or features within complex multi-aspect reviews. Without targeted aspect-level separation, the framework risks aggregating distinct sentiments into a single rating, potentially overlooking critical fine-grained insights essential for applications like product feedback or multi-feature analysis.

2.4 Methods Based on Recursive Neural Network

Recursive Neural Networks have gained traction in ABSA due to their strength in capturing the compositional structure of language, making them well-suited for modelling syntactic relationships. As discussed in the following subsections, their integration with recurrent architectures has opened promising avenues for combining hierarchical and sequential information in aspect-level sentiment modelling.

2.4.1 Recursive Neural Network

Recursive Neural Networks have emerged as powerful tools in aspect-based sentiment analysis due to their ability to capture hierarchical structures and compositional semantics within text [66]. Unlike recurrent models that process sequences linearly,

RecNNs operate over syntactic parse trees, enabling them to model how smaller linguistic units combine to form higher-level meanings. This makes them particularly effective in handling complex sentence constructions, such as negations, clauses, and nested opinions, which are often critical for correctly identifying sentiment at the aspect level. Recent research has explored combining RecNNs with other architectures, such as RNNs or attention mechanisms, to leverage both syntactic and sequential information, paving the way for more robust and fine-grained sentiment analysis models.

2.4.2 Combination of Recursive and Recurrent Neural Networks for Aspect-Based Sentiment Analysis

The study by Aydın and Güngör [66] introduces a novel ensemble framework that integrates Recursive Neural Networks and Recurrent Neural Networks to enhance aspect-based sentiment analysis. Recognizing that RNNs, particularly Gated Recurrent Units, are strong at modelling temporal sequences but lack syntactic understanding, the authors incorporated RecNNs to capture grammatical structures by parsing reviews into syntactic trees. By breaking down reviews into smaller subreviews centred on specific aspect terms and training recursive models on these structures, the framework ensures that each aspect's sentiment is accurately extracted before being passed to the recurrent layer. This dual approach allows the system to utilize both sequence information and tree-based hierarchical sentiment cues.

The research leveraged datasets from the SemEval-2014 Task 4 challenge, focusing on the restaurant and laptop domains. Experimental results showed that this combined model outperformed standalone RNNs and RecNNs, demonstrating a 1.6% improvement on average in classification accuracy. The recursive component was particularly effective when rooted embeddings were used, which encapsulated the entire sentiment flow of the sub reviews. Furthermore, the use of both constituency and dependency parsers enabled the model to handle complex linguistic phenomena such as negations and contrastive conjunctions. The ensemble method was shown to generalize well across both single-aspect and multi-aspect reviews, confirming its robustness in diverse evaluation settings.

Despite its strengths, the study exposes a remaining challenge: the difficulty of harmonizing the syntactic information from recursive modules with the temporal

dynamics modelled by recurrent layers. Specifically, inconsistencies arise when combining gold-standard labels with sentiment lexicon scores, sometimes leading to sentiment ambiguities at the word level. This highlights the broader limitation of current hybrid neural architectures in seamlessly integrating hierarchical and sequential representations, especially when dealing with long-range dependencies and aspect interactions.

2.5 Advancements to Deep Learning Models

Recent advancements in deep learning have significantly enhanced aspect-based sentiment analysis by introducing models capable of learning contextual and semantic representations automatically. This section outlines several improved architectures—such as CNN-BiLSTM with attention, AF-LSTM, Sentic LSTM, and GRU-based hybrids.

2.5.1 Aspect-Level Sentiment Analysis Using CNN, Bi-LSTM, and Attention Mechanism

Aspect-based sentiment analysis has advanced considerably, largely driven by innovations in deep learning architectures [67]. Lin and Najafabadi [68] proposed a deep learning architecture that integrates Convolutional Neural Networks, Bidirectional Long Short-Term Memory (Bi-LSTM) networks, and attention mechanisms to enhance aspect-level sentiment analysis. Their work aimed to tackle the persistent challenge of modelling long-range dependencies within customer reviews, recognizing that traditional models often fail to capture sentiment relationships between words or phrases located far apart in a text. By incorporating Bi-LSTM layers, the model processes information in both forward and backward directions, allowing it to retain context from surrounding words and improving its sensitivity to sentiment cues dispersed across a sentence or document. The attention mechanism further refines this process by directing the model's focus toward key phrases that carry sentiment relevance for specific aspects. This combination of CNNs for local feature extraction, Bi-LSTM for sequential context, and attention for selective focus represents an important advancement in refining sentiment analysis at the aspect level.

Despite these innovations, the study highlighted ongoing limitations, particularly concerning the model's capacity to maintain context over long sequences. While the

attention mechanism successfully emphasizes critical words within a sentence, its effectiveness varied across different datasets and domains. This inconsistency was more pronounced when applied to multilingual or domain-specific texts, where linguistic variations and specialized vocabulary posed additional challenges. The Bi-LSTM component, although designed to handle sequential data, still faced difficulties in capturing intricate dependencies that span multiple sentences. Consequently, the model occasionally misclassified sentiments when opinions evolved gradually or when sentiment expressions extended across different parts of a review.

These findings underscore a central challenge relevant to this research: the loss of context in long sequences. Even with the combined strengths of CNNs, Bi-LSTMs, and attention mechanisms, the architecture struggled to sustain coherent sentiment tracking in complex, multi-sentence customer reviews. This limitation suggests that further architectural improvements may be necessary to enhance the model's ability to handle lengthy and complex feedback. Addressing this issue is critical for developing more robust sentiment analysis systems that can reliably interpret the evolving sentiments embedded within extended customer narratives.

2.5.2 Learning to Attend via Word-Aspect Associative Fusion for Aspect-Based Sentiment Analysis

The Aspect Fusion LSTM (AF-LSTM) model, introduced by Chen et al. [69], presents an innovative method for improving how neural networks process aspect information in sentiment analysis. Rather than relying on simple concatenation of aspect and word embeddings, AF-LSTM employs a more sophisticated mechanism that enhances the model's capacity to capture the relationship between aspect terms and surrounding words within a sentence. Central to this approach is an attention mechanism that incorporates circular convolution and circular correlation, allowing the model to compute deeper interactions between aspect targets and contextual words. By applying these advanced operations, AF-LSTM increases the model's flexibility and connectedness, making it more effective at interpreting the intricate dynamics between different parts of a phrase a critical step toward improving aspect-level sentiment classification.

The authors evaluated AF-LSTM using well-established benchmarks, including the SemEval 2014 Task 4 and SemEval 2015 Task 12 datasets. Specifically, the SemEval

2014 Laptops and Restaurants datasets were used for aspect term classification, while a merged dataset from SemEval 2014 and 2015 supported aspect category classification. Their findings showed that AF-LSTM improved performance over baseline models, particularly when attention was directed toward the target aspect by averaging contextual attention scores. This focus enabled the model to prioritize sentiment cues near the aspect term, improving the accuracy of sentiment classification for specific features in customer reviews.

However, despite these advancements, the study revealed that AF-LSTM still faced challenges in managing loss of context in long. While the model excelled at capturing local interactions, its effectiveness diminished in sentences containing multiple aspects or where sentiment clues were distributed across extended passages. This limitation underscores the need for further improvements to extend the model's contextual memory and sustain sentiment tracking across longer sequences.

2.5.3 Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM

Researchers in the study by Zhang et al. [70] proposed an enhanced LSTM-based framework that integrates commonsense knowledge into a hierarchical attention mechanism for targeted aspect-based sentiment analysis. Their approach aimed to improve the model's ability to capture sentiments directed toward specific aspects within a review by combining both target-level and sentence-level attention layers. To further enrich the semantic representation of input text, they introduced Sentic LSTM, a variant of the traditional LSTM that allows commonsense knowledge to be embedded into the recurrent encoder. This integration provided the model with external conceptual understanding, enabling it to better infer implicit sentiment connections that might not be explicitly stated in the text.

The model was evaluated on two publicly available benchmark datasets, where it consistently outperformed baseline approaches in sentiment classification accuracy for targeted aspects. The hierarchical attention structure played a critical role in guiding the model to focus not only on words surrounding the target aspect but also on the broader sentence-level context. By leveraging both commonsense information and attention-based weighting, the model demonstrated improved capability in

identifying nuanced sentiment expressions related to different product features within a single review.

Despite these improvements, the study acknowledged a persistent limitation linked to a lack of aspect-specific focus. The aspect-level attention mechanism, although effective in many cases, proved difficult to train because it primarily relied on self-attention without adequately modelling the dependencies and interactions between multiple aspects within the same sentence. This shortcoming increased the risk of overlooking important sentiment cues tied to related aspects, especially in reviews discussing several features at once.

2.5.4 Aspect-Based Sentiment Analysis Using Attention Mechanism and Gated Recurrent Network

The study by Prabha et al. proposed an innovative approach for aspect-based sentiment analysis by integrating bidirectional Long Short-Term Memory, multiple attention mechanisms, and Gated Recurrent Units to address limitations in previous models [71]. Their architecture aimed to overcome the challenge of capturing sentiments when the target aspect and its sentiment expressions are not in close proximity within a sentence a common difficulty in real-world reviews. By leveraging position-weighted memory, the model assigned higher importance to sentiment words near the aspect, while applying multiple attention mechanisms to focus on the most relevant context. The output from the attention layers was non-linearly combined through a GRU, enabling the model to effectively aggregate critical information before sentiment classification via a softmax layer.

The experimental evaluation was conducted on the SemEval 2014 restaurant review dataset, where the proposed model achieved an accuracy of 81.23%, outperforming baseline methods like Support Vector Machine (80.16%), ATAE-LSTM (77.2%), and MemNet (80.95%). This superior performance was attributed to the combination of multiple attention mechanisms and GRU, which enhanced the model's ability to handle complex sentences with multiple aspects and polarities. The study further demonstrated the benefits of fine-tuning hyperparameters such as batch size, learning rate, and number of attention hops, showing improvements in predictive accuracy with higher epoch counts and optimized configurations.

Despite its notable advancements, the study highlighted that the model still inherits some limitations, particularly in handling more granular sentiment classifications beyond the traditional positive, neutral, and negative labels. Additionally, the framework relies heavily on predefined hyperparameter tuning, which may limit adaptability across diverse datasets and domains.

2.5.5 Positionless Aspect-Based Sentiment Analysis Using Attention

In their study, Yadav et al. (2021) proposed a novel approach to aspect-based sentiment analysis that eliminates the need for traditional positional embeddings, which are commonly used to represent the relationship between aspect words and their context [72]. The authors designed a simpler preprocessing method by replacing sentiment-carrying words using an opinion lexicon and masking the aspect word with a common token. This approach reduced the model's complexity while preserving essential positional information through pattern learning. They implemented a dual-input architecture using two Bidirectional GRU networks with attention layers: one for the original sentence embedding and another for the masked aspect embedding. Their experiments, conducted on standard SemEval datasets, demonstrated that the model achieved competitive or superior accuracy compared to state-of-the-art models that relied on complex positional embeddings and architectures.

Despite the innovation of using masking and lexicon replacement to implicitly capture positional cues, the model still faces challenges in accurately distinguishing neutral sentiments, as highlighted by their confusion matrix analysis. The system tends to misclassify neutral cases, likely due to its reliance on sentiment-bearing tokens like "positive" and "negative," which may overshadow more refined expressions of neutrality.

2.5.6 Hyperparameter Optimization in Deep Learning Models

Hyperparameter optimization plays a crucial role in enhancing the performance of deep learning models used in aspect-based sentiment analysis [73]. Every deep learning model depends on a set of hyperparameters that guide its learning process, including learning rate, batch size, number of hidden units, dropout rates, and sequence length. Finding the right combination of these hyperparameters can significantly influence the model's ability to generalize and make accurate predictions,

especially in complex tasks like ABSA, where multiple aspects and sentiments may coexist in a single review.

In research focused on ABSA, several approaches have been explored to optimize hyperparameters. A commonly used strategy is to adjust these values manually through trial and error, relying on the researcher's familiarity with the model and dataset [74]. Although this approach can sometimes produce satisfactory results, it is often laborious and may overlook better-performing configurations due to its reliance on subjective judgment and limited testing cycles. To address these shortcomings, more systematic approaches have been adopted in the literature, aiming to explore the hyperparameter space more thoroughly and efficiently.

Techniques such as grid search and random search have been widely applied in deep learning research to automate hyperparameter selection. Grid search tests all possible combinations within a predefined range, ensuring no configuration is missed within that grid, whereas random search selects random combinations, potentially covering broader areas of the hyperparameter space with fewer trials. Both approaches have been effectively used in sentiment analysis studies, offering structured ways to test different settings under controlled experimental conditions.

In recent studies, more advanced methods such as Bayesian optimization and gradient-based tuning have been employed to tackle the complexity of hyperparameter selection in deep learning models for ABSA. Bayesian optimization applies probabilistic models to prioritize testing of configurations that are likely to perform well, reducing unnecessary evaluations and saving computational resources. Gradient-based tuning treats certain hyperparameters as trainable parameters, allowing them to be adjusted dynamically during the training process [75]. These techniques have been credited for improving model performance without incurring the high computational cost associated with exhaustive searches.

Despite these advances, hyperparameter optimization remains a challenging aspect of deep learning for ABSA. Large models dealing with long reviews and multiple aspects often face intricate trade-offs between computational efficiency and predictive accuracy. Optimizing hyperparameters in such settings requires careful balancing to avoid overfitting to specific aspects or losing context in lengthy input sequences.

Researchers have increasingly recommended hybrid approaches that blend different optimization strategies or apply them selectively at various stages of model training to overcome these challenges.

2.6 Evaluation Metrics

Evaluation metrics play a vital role in the literature on aspect-based sentiment analysis by providing standardized methods for measuring model performance. Core metrics such as accuracy, precision, recall, and F1-score are commonly used to evaluate how well models classify aspect-level sentiments, while additional metrics including macro, micro, and weighted averages, confusion matrices, and aspect detection accuracy - offer deeper insights into classification reliability and model behaviour across varied sentiment classes.

2.6.1 Accuracy

Accuracy is one of the most fundamental and widely used evaluation metrics in machine learning and natural language processing tasks, including aspect-based sentiment analysis of customer reviews [81]. It measures the proportion of correct predictions made by a model out of all predictions. In the context of ABSA, accuracy quantifies how many aspect-level sentiment predictions (positive, negative, neutral) were correctly classified by the model relative to the total number of aspect-sentiment pairs evaluated [77]. This metric offers a straightforward assessment of a model's overall correctness in handling sentiment classification across multiple aspects.

Mathematically, accuracy is calculated by dividing the number of correctly predicted instances by the total number of predictions made. It can be expressed using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP = True Positives (correctly predicted positive sentiments)

TN = True Negatives (correctly predicted negative sentiments)

FP = False Positives (incorrectly predicted positive sentiments)

FN = False Negatives (incorrectly predicted negative sentiments)

In sentiment analysis research, accuracy has been extensively adopted as a baseline metric for evaluating classification models. Several studies in ABSA literature have reported accuracy scores to compare the effectiveness of various deep learning architectures. For instance, research utilizing recurrent neural networks, convolutional neural networks, and attention-based models consistently reported accuracy alongside other metrics to benchmark improvements in aspect-level sentiment classification.

2.6.2 Precision

Precision is a widely used evaluation metric in classification tasks, particularly in aspect-based sentiment analysis, where it measures the correctness of positive predictions. Specifically, precision quantifies the proportion of true positive predictions out of all instances that the model predicted as positive [77]. In the context of ABSA, precision reflects how accurately the model identifies positive sentiments for a given aspect without mistakenly labelling neutral or negative sentiments as positive. High precision indicates that the model makes few false positive errors, which is particularly important in applications where incorrectly classifying sentiment as positive could lead to misleading interpretations of customer feedback.

The mathematical formula for precision is defined as follows:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Where:

TP = True Positives (correctly predicted positive sentiments)

FP = False Positives (incorrectly predicted positive sentiments)

In sentiment analysis literature, precision has been extensively adopted to evaluate the accuracy of models in detecting specific sentiment classes. Several studies on ABSA have reported precision alongside recall and F1-score to provide a balanced evaluation of model performance [77], [82]. Researchers have used precision to assess how well deep learning models such as convolutional neural networks, recurrent neural networks, and attention-based architectures minimize false positive classifications when predicting aspect-level sentiment. Precision is often analysed in conjunction

with recall to capture both the selectivity and completeness of a model's predictions, allowing for a more refined interpretation of classification outcomes.

2.6.3 Recall

Recall is an essential evaluation metric in classification tasks, particularly in aspect-based sentiment analysis, as it measures the model's ability to correctly identify all relevant instances of a sentiment class [77]. In the context of ABSA, recall represents the proportion of true positive predictions out of all actual positive instances for a given aspect. A high recall score indicates that the model successfully captured most of the positive sentiments expressed toward an aspect, minimizing false negatives. This metric is critical in applications were failing to detect positive sentiments could result in overlooked customer satisfaction insights or missed opportunities for product improvement.

The mathematical expression for recall is given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Where:

TP = True Positives (correctly predicted positive sentiments)

FN = False Negatives (positive sentiments incorrectly classified as negative or neutral)

In ABSA literature, recall has been frequently reported alongside precision to evaluate the completeness of a model's sentiment classification [81]. Researchers have used recall to assess how effectively deep learning models, including RNNs, CNNs, and attention-based frameworks, identify all relevant sentiments associated with different aspects in customer reviews. This metric plays a pivotal role in determining the sensitivity of a model, particularly in scenarios involving multiple aspects with varying sentiment polarities, where missing a sentiment can distort the overall interpretation of customer opinions.

2.6.4 F1-Score

F1-Score is a widely adopted evaluation metric in classification tasks, particularly valuable in aspect-based sentiment analysis, as it provides a balance between precision

and recall. It is defined as the harmonic mean of precision and recall, offering a single metric that reflects both the correctness of positive predictions and the model's ability to capture all relevant positive instances [77]. In ABSA, F1-Score quantifies how effectively the model classifies positive sentiments for specific aspects while minimizing both false positives and false negatives. This metric is especially useful in imbalanced datasets, where high precision or recall alone may not adequately reflect a model's performance. The mathematical formulation for F1-Score is expressed as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In sentiment analysis research, F1-Score has been extensively reported as a key evaluation metric to assess deep learning models. Several studies have highlighted its relevance in capturing the trade-off between precision and recall, particularly in aspect-level sentiment classification tasks [83]. Researchers have used F1-Score to benchmark models such as convolutional neural networks, recurrent neural networks and attention-based architectures, enabling a balanced interpretation of model effectiveness across different sentiment classes. The metric has become standard practice in ABSA evaluations, offering a concise yet comprehensive measure of classification quality.

2.6.5 Macro-Averaged Precision, Recall, and F1-Score

Macro-averaged precision, recall, and F1-Score are evaluation metrics designed to provide an aggregate view of model performance across multiple classes by treating each class equally. In aspect-based sentiment analysis, these metrics are particularly useful when analysing sentiment across several categories such as positive, negative, and neutral, as well as across different product aspects [78]. Unlike micro-averaged metrics that weigh classes according to their frequency, macro-averaging computes precision, recall, and F1-Score independently for each class and then takes the unweighted mean across all classes. This approach ensures that minority classes are given equal importance in evaluating model performance, which is essential in sentiment analysis tasks with imbalanced class distributions. The mathematical expressions for macro-averaged metrics are defined as follows:

$$\text{Macro-Precision} = \frac{1}{N}\sum_{i=1}^{N} \text{Precision}_i$$

$$\text{Macro-Recall} = \frac{1}{N}\sum_{i=1}^{N} \text{Recall}_i$$

$$\text{Macro-F1} = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

Where:

N = total number of classes

Precision_i, Recall_i, $F1_i$ = precision, recall, and F1-Score for class i

In ABSA literature, macro-averaged metrics have been widely reported to evaluate sentiment classification models across all sentiment classes without bias toward more frequent categories [78]. Researchers have utilized these metrics to ensure that deep learning models achieve balanced performance, particularly in datasets where certain sentiment categories are underrepresented. By emphasizing equal treatment of all classes, macro-averaged precision, recall, and F1-Score serve as critical indicators of a model's generalization across diverse sentiment expressions and aspect categories.

2.6.6 Micro-Averaged Precision, Recall, and F1-Score

Micro-averaged precision, recall, and F1-Score are evaluation metrics that calculate global performance by aggregating contributions from all classes before computing the final metric. In aspect-based sentiment analysis, these metrics are particularly valuable when the dataset contains an imbalance in class distribution, as they assign greater influence to classes with more samples. Unlike macro-averaging, which treats each class equally, micro-averaging ensures that the evaluation reflects the model's performance across all predictions regardless of class frequency [78]. This makes micro-averaged metrics suitable for tasks where the majority class dominates or when the evaluation needs to focus on the overall number of correct predictions.

The mathematical expressions for micro-averaged precision, recall, and F1-Score are given by:

$$\text{Micro-Precision} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FP_i}$$

$$\text{Micro-Recall} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + \sum_{i=1}^{N} FN_i}$$

$$\text{Micro-F1} = 2 \times \frac{\text{Micro-Precision} \times \text{Micro-Recall}}{\text{Micro-Precision} + \text{Micro-Recall}}$$

Where:

N = total number of classes

TP_i, FP_i, FN_i = true positives, false positives, and false negatives for class i

In ABSA literature, micro-averaged metrics have been reported in studies focusing on large-scale sentiment datasets, where evaluating the aggregate predictive power across all classes is critical [78]. Researchers have applied micro-averaged precision, recall, and F1-Score to compare deep learning models in scenarios where accurate classification of the majority sentiment class remains a priority, while still accounting for minority class predictions within the total evaluation. These metrics have proven useful in practical applications where the overall number of correct classifications is more relevant than performance per individual class.

2.6.7 Weighted-Averaged Precision, Recall, and F1-Score

Weighted-averaged precision, recall, and F1-Score are evaluation metrics designed to provide a balanced assessment of model performance by accounting for the relative frequency of each class. In aspect-based sentiment analysis, these metrics are particularly useful when the dataset exhibits class imbalance, as they assign proportional weight to each class based on the number of true instances it contains. Unlike macro-averaging, which treats all classes equally, and micro-averaging, which emphasizes the total number of correct predictions, weighted-averaging strikes a middle ground by emphasizing the impact of both class size and predictive accuracy [78]. This ensures that the performance on majority classes contributes more heavily to the overall metric without completely ignoring minority classes.

The mathematical definitions for weighted-averaged precision, recall, and F1-Score are as follows:

$$\text{Weighted-Precision} = \frac{\sum_{i=1}^{N} w_i \times \text{Precision}_i}{\sum_{i=1}^{N} w_i}$$

$$\text{Weighted-Recall} = \frac{\sum_{i=1}^{N} w_i \times \text{Recall}_i}{\sum_{i=1}^{N} w_i}$$

$$\text{Weighted-F1} = \frac{\sum_{i=1}^{N} w_i \times F1_i}{\sum_{i=1}^{N} w_i}$$

Where:

N = total number of classes

w_i = number of true instances for class i

Precision_i, Recall_i, $F1_i$ = precision, recall, and F1-Score for class i

In ABSA research, weighted-averaged metrics have been employed to ensure a fair evaluation of models trained on imbalanced datasets, such as those where certain aspects or sentiment classes are overrepresented [78]. Researchers have used these metrics to evaluate deep learning models, including RNNs, CNNs, and attention-based architectures, highlighting their usefulness in reflecting both predictive performance and real-world class distributions. Weighted-averaged precision, recall, and F1-Score are particularly valued in industry applications where misclassifying majority class instances may carry higher practical implications than minority classes, yet overall fairness remains important.

2.6.8 Confusion Matrix

The confusion matrix is a widely used evaluation tool in classification tasks, providing a detailed breakdown of a model's predictive performance by displaying the number of correct and incorrect predictions across all sentiment classes. In aspect-based sentiment analysis, the confusion matrix offers insight into how well the model distinguishes between positive, negative, and neutral sentiments for different aspects. The matrix is structured as a table with rows representing the actual classes and columns representing the predicted classes, allowing researchers to identify specific patterns of misclassification across sentiment categories.

In ABSA literature, confusion matrices have been used to analyze how deep learning models handle misclassifications between closely related sentiment categories, offering insights into where models tend to confuse neutral with positive or negative sentiments [79]. Researchers have leveraged this matrix to pinpoint specific challenges in aspect-level sentiment classification, especially in datasets where sentiments are subtle or expressed indirectly. By visualizing classification outcomes, the confusion matrix supports a more granular evaluation beyond aggregated metrics like accuracy or F1-Score.

However, despite its usefulness, the confusion matrix presents limitations in large-scale sentiment analysis. One major challenge is that it becomes cumbersome and harder to interpret as the number of sentiment classes or aspect categories increases. This complexity may obscure actionable insights when models are evaluated across multiple aspects and fine-grained sentiment labels [84]. Additionally, the confusion matrix does not provide a single, summarized performance score, requiring additional metrics to interpret the overall effectiveness of a model [85]. These limitations suggest that while the confusion matrix is valuable for detailed error analysis, it needs to be complemented with other evaluation metrics to achieve a comprehensive assessment of ABSA model performance.

2.6.9 Aspect Detection Accuracy

Aspect Detection Accuracy is a specialized evaluation metric used in aspect-based sentiment analysis to measure the model's ability to correctly identify and classify the relevant aspects mentioned within a review [80]. In ABSA tasks, sentiment classification alone is insufficient without accurately detecting the specific product features or aspects to which sentiments are directed. Aspect Detection Accuracy therefore quantifies the proportion of aspects that the model correctly identifies out of all the true aspects present in the text. A higher score indicates the model's competence in recognizing aspect terms, which is critical for generating fine-grained sentiment insights in multi-aspect reviews.

The mathematical formulation for Aspect Detection Accuracy is expressed as:

$$\text{Aspect Detection Accuracy} = \frac{\text{Number of correctly detected aspects}}{\text{Total number of actual aspects}}$$

In the ABSA literature, Aspect Detection Accuracy has been applied as a key metric to evaluate models that integrate aspect extraction with sentiment classification, particularly in deep learning frameworks that combine recurrent or convolutional architectures with attention mechanisms [80]. Researchers have used this metric to benchmark models across different product domains, such as restaurant and laptop reviews, ensuring that sentiment predictions are linked to the correct aspects. Accurate aspect detection is fundamental for enabling actionable insights, as misidentifying aspects can lead to misinterpretation of customer feedback or incorrect product improvement priorities.

Despite its usefulness, Aspect Detection Accuracy presents notable limitations. One key challenge is that the metric only evaluates whether an aspect was identified or missed, without considering partial matches or semantic similarity. This binary evaluation may penalize models for near-correct predictions where synonymous or paraphrased aspect terms are used. Additionally, in reviews with implicit aspects or ambiguous expressions, aspect detection becomes more difficult, leading to underestimation of a model's actual capability.

2.7 Summary

This literature review highlights critical gaps in current deep learning approaches to aspect-based sentiment analysis. Existing models such as CNNs, RNNs, and RecNNs often struggle with long-range dependencies, especially when handling multi-sentence reviews. This limits their ability to retain contextual coherence and accurately classify sentiments linked to distant or multiple aspects. Additionally, many models lack fine-grained aspect-specific focus, treating all words equally and overlooking sentiment cues tied to specific features. Even those incorporating attention mechanisms frequently fall short in managing overlapping sentiments or resolving conflicts in opinion-rich content. The evaluation metrics widely used in ABSA may not fully capture aspect-level errors. They offer a general performance view without addressing deeper issues such as aspect ambiguity, or explainability.

To address these gaps, the proposed study introduces key architectural innovations, including the integration of multi-head attention mechanisms, dual attention mechanisms, and the optimization of hyperparameters. These enhancements aim to improve sentiment discrimination across complex sentence structures and enhance the

model's ability to isolate and classify sentiments tied to specific aspects with higher accuracy.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

Chapter Three outlines the methodology employed in this research, systematically describing the steps undertaken to achieve the study's objectives. This chapter provides a structured explanation of the research design that was adopted by the researcher and illustrates the research process that the researcher undertook in achieving the objectives of this study. Data acquisition for this study is also described in this chapter, alongside the procedure followed in developing and evaluating the improved model. Finally, this chapter describes the experimental materials used by the researcher and the ethical considerations that underpin this study.

3.2 Research Design

This study adopted a true experimental research design which is characterized by its capacity to manipulate independent variables, control extraneous variables, and randomly assign subjects to different experimental conditions [86]. Recognized as a robust approach for establishing cause-and-effect relationships, this design offers a systematic framework for isolating the impact of specific interventions on measured outcomes. Among the different types of true experimental designs, the Pretest-Posttest Control Group Design was selected for this study, given its suitability in comparing the improved model against the baseline model under controlled conditions. The decision to use this research design was justified by the level of control it afforded throughout the research process. The independent variables could be manipulated, enabling the researcher to observe their direct effects on the dependent variables [87].

Participants were randomly assigned into two groups as part of this design: an experimental group, which interacted with the improved ABSA model incorporating attention mechanisms and other architectural enhancements, and a control group, which relied on a standard baseline model without such improvements. Both groups participated in a pretest and a post-test to assess their performance. The pretest measured the models' initial performance on a standardized dataset prior to the application of any interventions, whereas the post-test evaluated performance following the implementation of the model enhancements. A comparison of pretest and post-test results between the experimental and control groups enabled a rigorous

evaluation of the model's effectiveness in improving aspect-level sentiment classification. This design ultimately facilitated the establishment of a causal relationship between the model improvements and the observed improvements in performance metrics.

3.3 Research Process

The research process in this study detailed the sequential steps undertaken to accomplish the research objectives and provide answers to the stated research questions [86]. The general objective of the study was divided into three specific objectives, each of which was addressed systematically to ensure that the overall aim of improving aspect-based sentiment analysis through deep learning enhancements could be achieved. Each objective served as a milestone leading toward the development and evaluation of the improved ABSA model.

The first specific objective focused on determining architectural features that can improve the performance of a suitable deep learning for aspect-based sentiment analysis. To address this objective, a comparative analysis of several baseline deep learning models was conducted, as detailed in section 3.5, which is a standard practice in this area of research in identifying baseline models. The models were subjected to identical experimental conditions, allowing for a fair comparison of their accuracy, precision, recall, and F1 score on standardized datasets. The outcome of this stage was the selection of the most suitable model architecture and a set of performance-enhancing features for further development.

The second specific objective was to develop an improved deep learning model incorporating attention mechanisms tailored for aspect-based sentiment analysis of customer reviews. The improvements aimed to address key limitations outlined in the problem statement and literature review, including challenges such as loss of context in long sequences, lack of aspect-specific focus, and difficulty handling multi-aspect sentiments. Section 3.6 provides a comprehensive explanation of how the improved model was developed, incorporating enhancements such as attention mechanisms, encoder-decoder structures, and optimized hyperparameters. The deliverable from this phase was an improved model.

The third specific objective was to evaluate the performance of the improved model on the aspect-based sentiment analysis benchmark datasets. This evaluation was conducted as described in section 3.7, involving the evaluation of the improved model on benchmark datasets and the measurement of its performance across standard evaluation metrics, namely; Accuracy, precision, recall, F1-score. Figure 3.1 illustrates the sequential flow of the research process, showing how each specific objective contributed to the achievement of the general research objective.

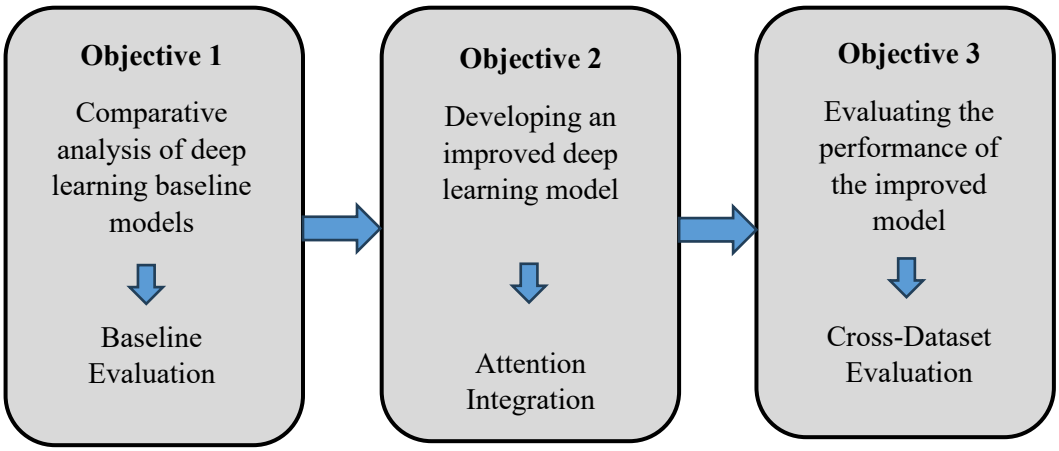


Figure 3.1: The Research Process

3.4 Data Acquisition

This study utilized secondary data from publicly available benchmark datasets tailored for aspect-based sentiment analysis, following the precedent set by prior research [88], [89], [90], [91]. These datasets offered standardized, accessible, and diverse review samples, making them ideal for training and evaluating deep learning models. Given the complexity and time demands of aspect-level annotation, using secondary data also allowed the study to focus on model development and performance assessment efficiently. The target population consisted of customer reviews from the SemEval Task 4 and Multi-Aspect Multi-Sentiment (MAMS) datasets.

3.4.1 Dataset Description

This study utilized two publicly available benchmark datasets commonly used in aspect-based sentiment analysis research: the SemEval Task 4 dataset and the Multi-Aspect Multi-Sentiment dataset. The SemEval dataset was selected as the training and validation set, while the MAMS dataset was used to evaluate how the proposed model would generalize on data it had not been trained on. Both datasets were chosen due to

their relevance, wide acceptance in natural language processing research, and suitability for evaluating aspect-level sentiment classification models. Their adoption was further supported by their high-quality annotations, standardized structure, and prior use in similar studies [92], [93], facilitating comparability and reproducibility of results.

The SemEval Task 4 dataset, publicly released by the International Workshop on Semantic Evaluation, is a well-established benchmark for aspect-based sentiment analysis. It comprises two domains: Restaurant reviews and Laptop reviews, each annotated with sentiment polarities (positive, negative, neutral) at the aspect level. This dataset has been extensively employed in ABSA research to evaluate deep learning models, providing a standardized framework for performance comparison [92], [93]. The dataset contains separate training and testing sets for both domains, with balanced representation across sentiment classes, making it an ideal choice for training and validating sentiment classification models. Table 3.1 presents a distribution of the SemEval Task 4 dataset, summarizing the number of instances across sentiment classes split in the ration 70:20:10.

Table 3.1: SemEval Task 4 Dataset Distribution

Dataset	**Positive**			**Neutral**			**Negative**		
	Train Set	Validation Set	Test Set	Train Set	Validation Set	Test Set	Train Set	Validation Set	Test Set
Restaurant	2025	578	289	583	167	83	702	201	100
Laptop	934	267	134	443	127	63	698	200	100
Total	2959	845	423	1026	294	146	1400	400	200

The MAMS dataset was designed to reflect multi-aspect, multi-sentiment scenarios, offering a challenging benchmark for evaluating models that must classify sentiments across multiple aspects within a single review. It contains two primary subsets: Aspect-Term Sentiment Analysis (ATSA) and Aspect-Category Sentiment Analysis (ACSA). The ATSA subset focuses on sentiment classification at the aspect term level, while the ACSA subset evaluates sentiment at a broader aspect category level. The dataset includes reviews covering diverse product features, with many sentences

expressing multiple sentiments across different aspects. Table 3.2 presents a distribution of the MAMS dataset.

Table 3.2: Multi-Aspect Multi-Sentiment Dataset

Dataset		Positive	Negative	Neutral	Total
ATSA	Train	3380	2764	5042	11186
	Train (small)	1089	892	1627	3608
	Validation	403	325	604	1332
	Test	400	329	607	1336
ACSA	Train	1929	2084	3077	7090
	Train (small)	1000	1100	1613	3713
	Validation	241	259	388	888
	Test	245	263	393	901

3.4.2 Data Pre-processing

Data preprocessing was carried out on both the SemEval Task 4 and MAMS datasets to ensure the textual data was in a suitable format for training and evaluating deep learning models for aspect-based sentiment analysis. This step was essential for preparing the data to meet the input requirements of the model and to enhance the consistency and quality of the input sequences [94]. Preprocessing in ABSA plays a critical role not only in standardizing the input but also in preserving aspect-level sentiment information, which is vital for achieving accurate sentiment classification across multiple aspects within a single review.

The initial stage of preprocessing involved aspect-level text segmentation and tokenization. Each review sentence was segmented into tokens, with special attention to preserving aspect terms and their surrounding context. This ensured that the tokenized sequence retained meaningful relationships between aspect terms and sentiment-bearing words [95]. The tokenization process aimed at converting the segmented text into numerical token indices, along with the generation of attention masks to guide the model in identifying which tokens were contextually relevant for aspect-level sentiment analysis.

The next stage addressed the challenge of handling variable-length sentences across the datasets. Since the review sentences varied in length, sequences longer than the

defined maximum input length were truncated strategically to retain aspect terms and their immediate context, minimizing the risk of losing important sentiment clues. Sequences shorter than the required length were padded using a designated padding token to achieve uniform input lengths. This approach enabled the model to process input sequences in fixed-length batches, facilitating efficient training without compromising the alignment of aspect-related sentiment information.

Following tokenization, truncation, and padding, the data was split into training, validation, and testing sets using a 70:20:10 ratio, a practice commonly adopted in machine learning to balance model training and evaluation. The training set was used to fit the model's parameters, the validation set provided checkpoints for monitoring performance and tuning hyperparameters, and the test set evaluated the model's generalization capability on unseen data. Through these preprocessing steps, the datasets were transformed into structured, model-ready input while preserving the aspect-specific sentiment details critical for aspect-based sentiment analysis.

3.5 A Comparative Study of Existing Deep Learning Models

A comparative study was conducted to evaluate the performance of various deep learning models for aspect-based sentiment analysis, both in their baseline form and with integrated attention mechanisms. The study examined four baseline architectures commonly used in sentiment analysis: Recurrent Neural Networks, Convolutional Neural Networks, Recursive Neural Networks, and Memory Networks [26], [49], [96], [97]. Each baseline model was tested under two experimental settings: without any attention mechanism and with attention mechanisms incorporated.

To comprehensively assess the influence of attention mechanisms on model performance, four different attention strategies self-attention, multi-head attention, global attention, and hierarchical attention were individually applied to each baseline model. This design resulted in multiple experimental variations, enabling a systematic evaluation of how each attention type enhanced or influenced sentiment classification within each architecture.

All models were trained and evaluated under identical experimental conditions to ensure fairness and comparability. The experiments were conducted using the SemEval Task 4 dataset, following the preprocessing procedure described in Section

3.4.2. The dataset was divided into training, validation, and test sets in a 70:20:10 ratio. Model training was performed in a controlled computational environment using Google Collaboratory Pro, equipped with GPU acceleration, PyTorch version 2.5, and CUDA version 12.6. To maintain consistency, all models were trained using the same hyperparameters: a learning rate of 1e-3, a batch size of 64, ten epochs, and the Adam optimizer with a cross-entropy loss function.

Each baseline deep learning model was trained five times: once in its original form without attention, and four additional times, each time paired with one of the four attention mechanisms. For instance, the RNN was evaluated independently with self-attention, multi-head attention, global attention, and hierarchical attention; the same process was repeated for CNN, RecNN, and MN architectures. This structured experimentation allowed a direct comparison of how each attention mechanism affected model performance within each neural network type. The attention mechanisms were incorporated as additional layers interfacing with the respective model architectures, ensuring that their implementation followed established integration practices for each model type.

Each model was built with architecture-specific configurations. The RNN used a two-layer stacked structure with tanh activations, processing word embeddings sequentially before classification. The CNN applied 1D convolutional layers with ReLU activations and max-pooling, followed by a dense layer. The RecNN adopted a tree-based structure with ReLU activations for phrase-level sentiment. The Memory Network (MN) utilized memory slots with soft attention and a linear classification layer. All models included dropout regularization and were implemented in PyTorch under identical training conditions for fair comparison.

Evaluation of the models was based on accuracy, precision, recall, and F1-score to assess classification performance across positive, negative, and neutral sentiment categories for each detected aspect. The evaluation results for each combination of the baseline model and attention mechanism were compared to determine which configuration yielded the highest performance. The findings from this comparative study informed the selection of the best-performing model and attention mechanism pairing, which formed the basis for further improvements in the second objective of

the research. Detailed experimental results from this comparative analysis are presented and discussed in Chapter 4 of this thesis.

3.6 Developing the Improved Model

To address limitations in the baseline RNN+Hierarchical Attention architecture especially difficulties with multi-aspect reviews, aspect-specific focus, and context retention across long inputs - an improved model was developed. Enhancements included Multi-Head Hierarchical Attention, Dual Attention Mechanisms, and optimized hyperparameters to boost sentiment classification accuracy.

3.6.1 Multi-Head Hierarchical Attention Mechanism

One critical limitation identified in the baseline model was its inability to capture complex dependencies in customer reviews containing sentiments linked to multiple aspects. This problem, previously discussed in the literature and reinforced during baseline experimentation, resulted from the model's tendency to collapse sentiment representations across aspects, especially when reviews expressed mixed or overlapping opinions.

Baseline RNN architecture for aspect-based sentiment analysis initiates with an input review that undergoes preprocessing, including steps like tokenization and lemmatization. Cleaned text is then transformed into dense vector representations through the Glove word embedding technique capturing semantic relationships among words. These embeddings are passed into a stacked Recurrent Neural Network, which excels at modelling sequential data and preserving contextual dependencies across time steps. Output from the RNN is directed to a classification layer that predicts sentiment polarity - positive, negative, or neutral - for each detected aspect in the review. Finally, an aspect-sentiment pairing module maps each aspect term to its corresponding sentiment label, enabling sentiment interpretation across aspects. Figure 3.2 illustrates this Baseline RNN Architecture.

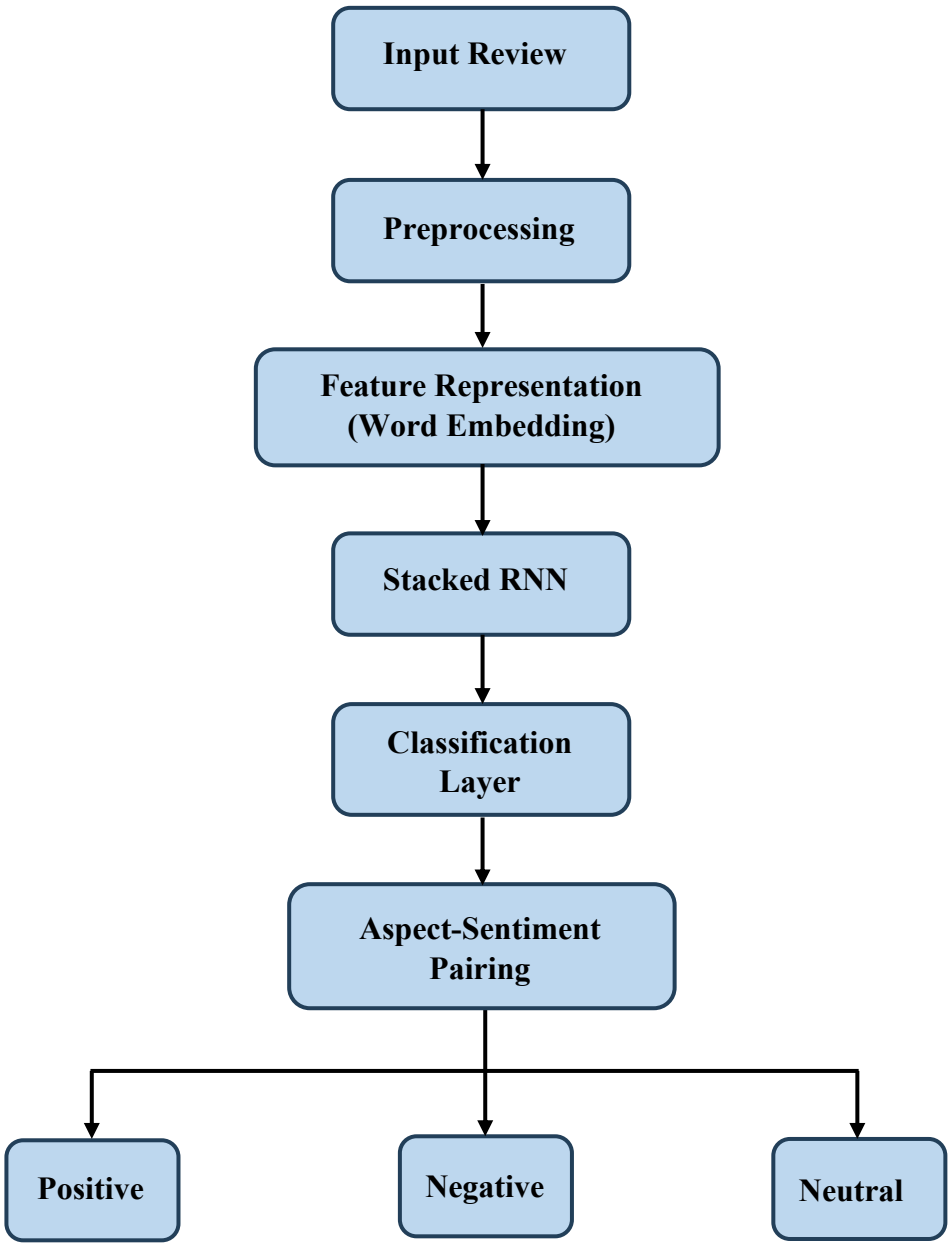


Figure 3.2: Baseline RNN Architecture

The addition of a hierarchical attention mechanism to the baseline improved the classification of multiple aspects. The input review undergoes preprocessing and word embedding before entering the stacked RNN. From there, attention is applied at two levels: word-level attention highlights important tokens, while sentence-level attention selects key sentences. The refined output passes through the classification layer to predict sentiment polarity. Finally, aspect-sentiment pairing links each aspect to its corresponding sentiment, as illustrated in Figure 3.3.

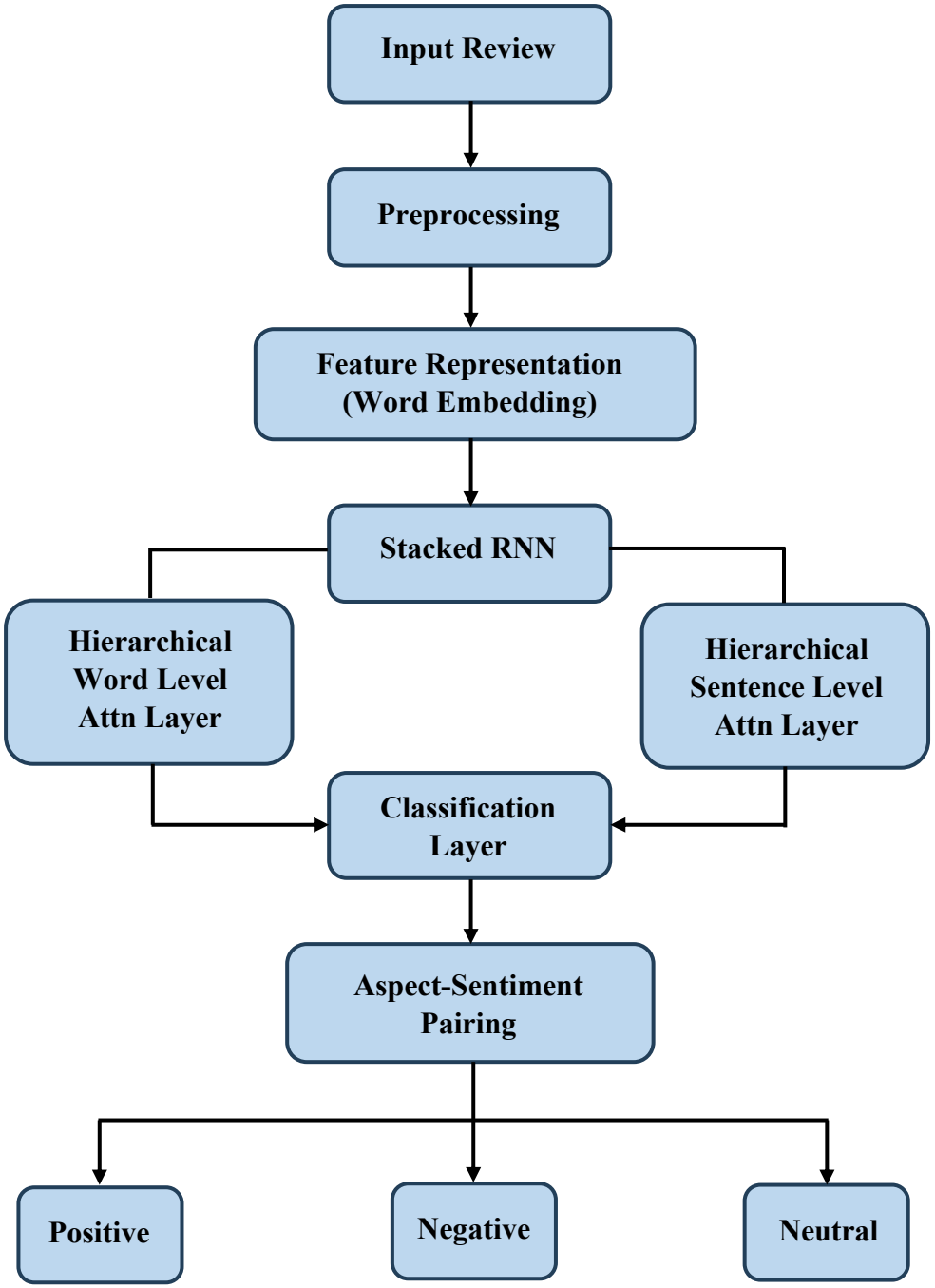


Figure 3.3: Baseline RNN with Hierarchical Attention

Attention-based models still struggle with aspect-based sentiment analysis, particularly when sentiments are distributed across different parts of a review or when multiple aspects are discussed simultaneously. To address this issue, the researcher extended the original RNN+Hierarchical Attention architecture by introducing a Multi-Head Hierarchical Attention Mechanism, customized for multi-aspect ABSA tasks.

The implementation involved two distinct levels of attention: word-level and sentence-level, each comprising multiple parallel attention heads. These heads were designed to focus on different semantic regions of the input text, allowing the model to extract multiple sentiment cues relevant to various aspects. Input reviews were first pre-processed and transformed into word embeddings using pre-trained GloVe vectors, then passed through a stacked RNN to capture sequential dependencies. The resulting hidden states were processed by the word-level multi-head attention layer, where each head independently computed attention weights to highlight sentiment-bearing tokens.

Outputs from all heads were concatenated and projected into a fixed-dimensional vector that encapsulated diverse sentiment signals. This was then passed into the sentence-level multi-head attention layer, which applied a similar multi-head attention process over the sentence representations. Each head contributed to understanding different opinion flows across sentence segments, which is crucial in reviews containing several aspect-opinion pairs. The final representation was directed to the classification layer, followed by an aspect-sentiment pairing module that assigned sentiment labels - positive, negative, or neutral to each aspect.

This architecture enhanced the model's ability to disentangle overlapping sentiment expressions and improved interpretability in multi-aspect scenarios. As a result, it resolved key limitations in previous models related to diluted focus and sentiment entanglement. Figure 3.4 illustrates the integrated Multi-Head Hierarchical Attention mechanism, showing how each attention layer enriches the model’s capacity to handle complex, aspect-rich customer reviews.

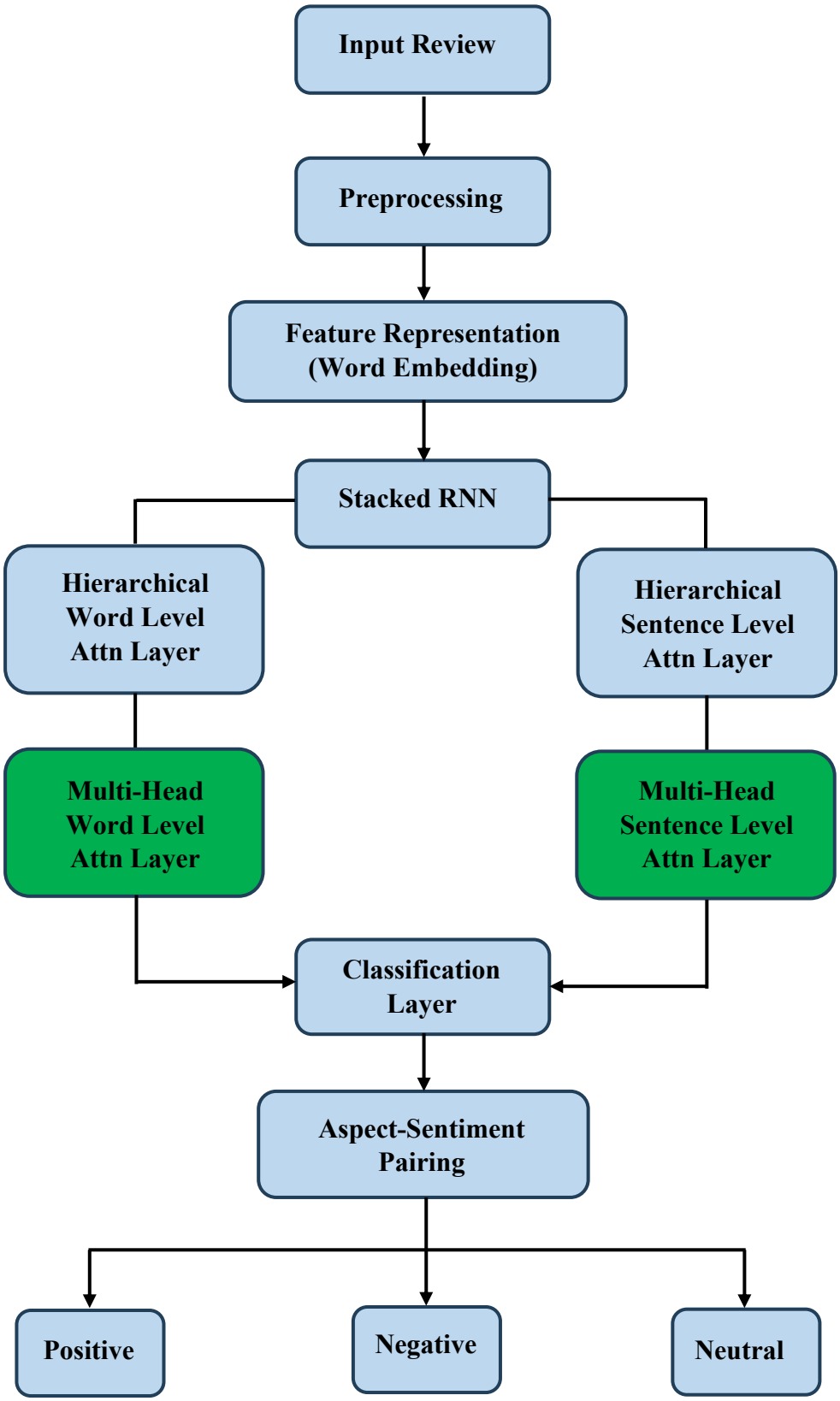


Figure 3.4: Integrated Multi-Head Attention Mechanism

The Multi-Head Hierarchical Attention Algorithm refines sentiment classification by applying attention mechanisms at both the word and sentence levels. It begins by embedding input tokens and processing them through stacked RNN layers, followed by multiple attention heads that extract salient word and sentence features. The final attention-weighted representation is then used to predict sentiment via a dense softmax layer. The detailed steps of the algorithm are presented below to illustrate the full processing flow from input to sentiment prediction.

Multi-Head Hierarchical Attention Algorithm

Require: Tokenized review sentence S = {w_1, w_2, ..., w_T},

aspect term A, number of heads H

Ensure: Final attention-weighted representation for sentiment classification

// Embed words, E = embed(S) // Generate word embeddings for each token

// Encode sequence with stacked RNN

H_seq = RNN_stack(E) // Hidden states from stacked RNN layers

// Apply word-level multi-head attention

for head in 1 to H **do**:

word_attn_weights[head] = attention_head(H_seq, head)

word_attn_output[head] = weighted_sum(H_seq, word_attn_weights[head])

end for

// Concatenate and project to fixed dimension,W_word = concat(word_attn_output[1:H])

P_word = project(W_word) // Dimensionality reduction (optional)

// Generate sentence vector

S_vec = aggregate(P_word) // May be average, max, or learned weighting

//Apply sentence-level multi-head attention

for head in 1 to H **do**:

sent_attn_weights[head] = attention_head(S_vec, head)

sent_attn_output[head] = weighted_sum(S_vec, sent_attn_weights[head])

end for

//Concatenate outputs and project

W_sent = concat(sent_attn_output[1:H]), final_representation = project(W_sent)

// Predict sentiment, sentiment_logits = softmax(dense(final_representation))

return sentiment_logits

Multi-head hierarchical attention approach offered a substantial improvement over the single-head attention mechanisms by enabling the model to process multiple sentiment expressions in parallel, each potentially linked to a different aspect. It resolved the issue of diluted sentiment focus observed in earlier models and allowed for more granular sentiment analysis, especially in reviews where opinions on various product features were distributed across different parts of the text. The implementation of multi-head attention at both hierarchical levels gave the model a richer, more structured understanding of the review, addressing the limitation of weak multi-aspect comprehension.

3.6.2 Dual Attention Mechanism for Aspect-Specific Focus

Another major limitation encountered in the baseline models was the lack of aspect-specific focus. The model's attention mechanisms though effective in highlighting salient tokens did not always distinguish which words contributed sentiment toward a specific aspect, especially in multi-aspect reviews. This often resulted in misclassification where sentiment-bearing expressions were incorrectly attributed to the wrong aspect. To overcome this limitation, a Dual Attention Mechanism was integrated into the baseline architecture. This mechanism explicitly modelled the bidirectional relationship between the aspect and its surrounding context, allowing the model to better align sentiment expressions with their intended aspect.

The dual attention mechanism operated through two interdependent attention flows: Aspect-to-Context Attention, where the given aspect term was used as a query to selectively attend to relevant context words in the review. Context-to-aspect attention, where the surrounding sentence context guided the model in refining its understanding of the aspect itself, particularly useful when the aspect term was ambiguous or general (e.g., "service" or "performance"). In the first stage, the aspect embedding was concatenated with each word-level hidden state from the RNN layer. An attention weight was computed by measuring the similarity between the aspect vector and each contextual word, producing a context vector weighted by aspect relevance. This vector represented the model's interpretation of the sentence, filtered through the lens of the target aspect.

In the second stage, the model reversed the process. A context summary vector, derived from the initial attention output, was used to guide the model in refining its

representation of the aspect. This ensured that the aspect term itself was interpreted relative to the full sentence, capturing implicit sentiment. The sentence-level and aspect-level vectors were fused using element-wise addition. The two attention outputs were then merged through concatenation and a non-linear transformation, producing a joint representation that fused both the aspect's interpretation and its relevant sentiment cues. This dual-flow architecture allowed the model to distinguish between overlapping and interacting aspects. For instance, in a review like “The battery drains fast but the screen is gorgeous,” the model learned to attend to “drains fast” only when analysing “battery”, and “gorgeous” only when evaluating “screen”. This addressed the core issue of aspect entanglement, where shared or nearby sentiment words previously misled single-path attention mechanisms.

The dual attention mechanism was implemented using standard dot-product attention for both directions, combined with gating layers to control the influence of each flow during training. This setup was compatible with the previously introduced multi-head hierarchical architecture and required only minimal adjustment to the existing training loop. Figure 3.5 shows the dual attention mechanism for aspect-specific focus.

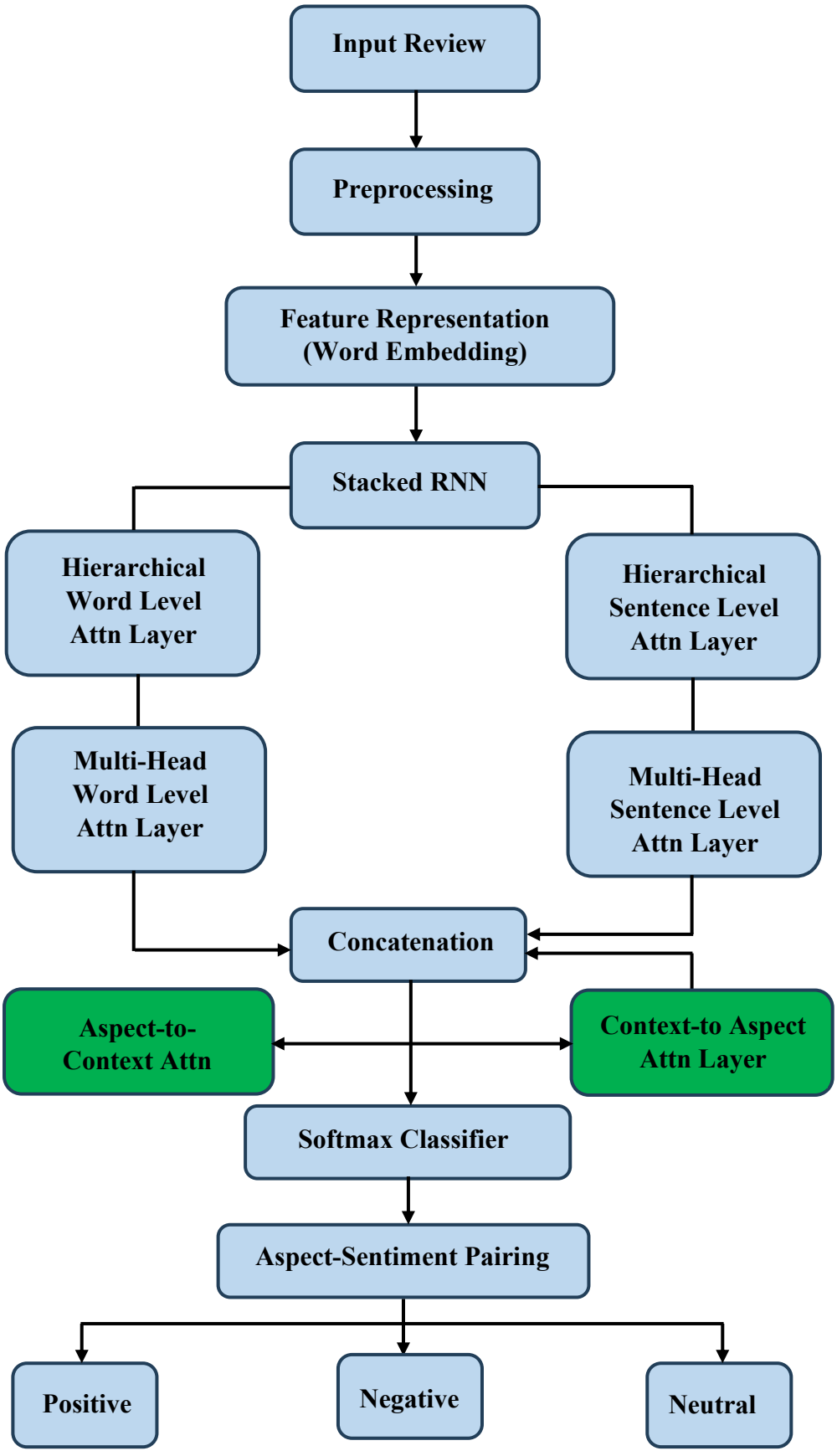


Figure 3.5: Dual Attention Mechanism for Aspect Specific Focus.

Dual Attention Mechanism Algorithm was designed to enhance the model's ability to focus simultaneously on both the aspect term and its surrounding context. It strengthened the alignment between aspect-specific semantics and relevant contextual cues, making it more effective for fine-grained sentiment classification. To operationalize this dual focus, the following Algorithm was developed.

Dual Attention Mechanism Algorithm

Require: Sentence tokens S = {w_1, w_2, ..., w_T}, aspect term A, number of hidden units H

Ensure: Final sentiment prediction based on aspect-specific dual attention

// Generate word embeddings for sentence and aspect

E_sentence = get_embeddings(S), E_aspect = get_embedding(A)

// Encode the sentence with stacked RNN

for layer_index from 1 to number_of_layers **do:**

if layer_index == 1 **then**: H_seq = RNN_layer(E_sentence)

else:

H_seq = RNN_layer(H_seq)

end if

end for

//Compute Aspect-to-Context Attention

for t from 1 to T **do:** attention_weight_ac[t] = dot_product(E_aspect, H_seq[t])

end for

normalized_ac = softmax(attention_weight_ac)

context_vector = weighted_sum(H_seq, normalized_ac)

//Compute Context-to-Aspect Attention, summary_context = average(H_seq)

attention_weight_ca = dot_product(summary_context, E_aspect)

normalized_ca = softmax(attention_weight_ca)

refined_aspect = weighted_sum(E_aspect, normalized_ca)

//Merge both attention vectors

combined_vector = concatenate(context_vector, refined_aspect)

projected_vector = ReLU(dense_layer(combined_vector))

//Predict sentiment, sentiment_logits = softmax(output_layer(projected_vector))

return sentiment_logits

3.6.3 Hyperparameter Optimization

Loss of critical contextual information in lengthy customer reviews was the third major limitation identified in the baseline model. This limitation was primarily caused by inadequate sequence representation, poor regularization, and suboptimal training dynamics, which resulted in instances where earlier or later sentiment-bearing clauses were overlooked due to restricted input lengths or vanishing gradients during model training. To mitigate this, the researcher conducted hyperparameter optimization geared towards increasing the model's ability to effectively preserve and utilize context over extended input spans.

Manual tuning combined with automated search-based strategies were adopted as the hyperparameter optimization approaches in this research. The rationale for using both was to strike a balance between domain-informed adjustments and systematic exploration of the model's hyperparameter search space. The focus was on hyperparameters that directly influence sequence length handling, training stability, and context retention, such as maximum input length, hidden layer dimensions, learning rate, dropout rate, number of RNN layers, batch size, and gradient clipping thresholds.

In the initial phase, the researcher manually adjusted static hyperparameters. In this context, such hyperparameters included the maximum sequence length, which was optimized by incrementally increasing the token limit to ensure that longer reviews were not aggressively truncated. Similarly, the dropout rate and number of RNN layers were adjusted iteratively while monitoring validation loss during model training and validation, enabling the model to learn deeper patterns without overfitting the training data. Each modification of these hyperparameters was evaluated using a validation-based checkpoint system to track performance improvements during training and validation of the model.

For the dynamic hyperparameters such as learning rate, weight decay, and optimizer configurations, an automated optimization loop was implemented using the Optuna framework. The researcher employed Bayesian optimization with adaptive pruning to identify the most suitable hyperparameter configurations based on validation loss trends. The hyperparameter search space Θ was initialized with domain-informed bounds. At each iteration t, a candidate tuple $\theta_t \in \Theta$ was sampled, and the model was

trained briefly to estimate validation performance. The update rule for each parameter was governed by the following gradient-based optimization principle in [74]:

$$[\theta_i^{(t+1)} = \theta_i^{(t)} - \eta \cdot \nabla_{\theta_i} \mathcal{L}]$$

Where:

θ_i is the hyperparameter being optimized

η is the learning rate or step size for updating

L is the validation loss

t is the current iteration

The optimization was terminated when no significant improvements in validation loss were observed across multiple trials, or when the validation curve stabilized. To further ensure stable training across long sequences, the researcher introduced gradient clipping to a fixed threshold, preventing exploding gradients from destabilizing backpropagation in deep RNN layers. A range of clip norms was evaluated with a final selection based on empirical stability during training epochs. This hybrid optimization approach allowed the model to better generalize on multi-sentence, multi-aspect reviews, especially where context was distributed across wide spans. The incorporation of tailored hyperparameter strategies addressed the root cause of performance degradation in long texts namely, model overfitting on short-term patterns and insufficient capacity to encode lengthy dependencies.

The final hyperparameter configuration yielded consistent validation improvements, reduced convergence time, and enhanced the robustness of the model when handling verbose or compound sentiment expressions. The full suite of optimal hyperparameters derived from this process is presented in Chapter 4. The hyperparameter optimization process adopted is summarized in Figure 3.6.

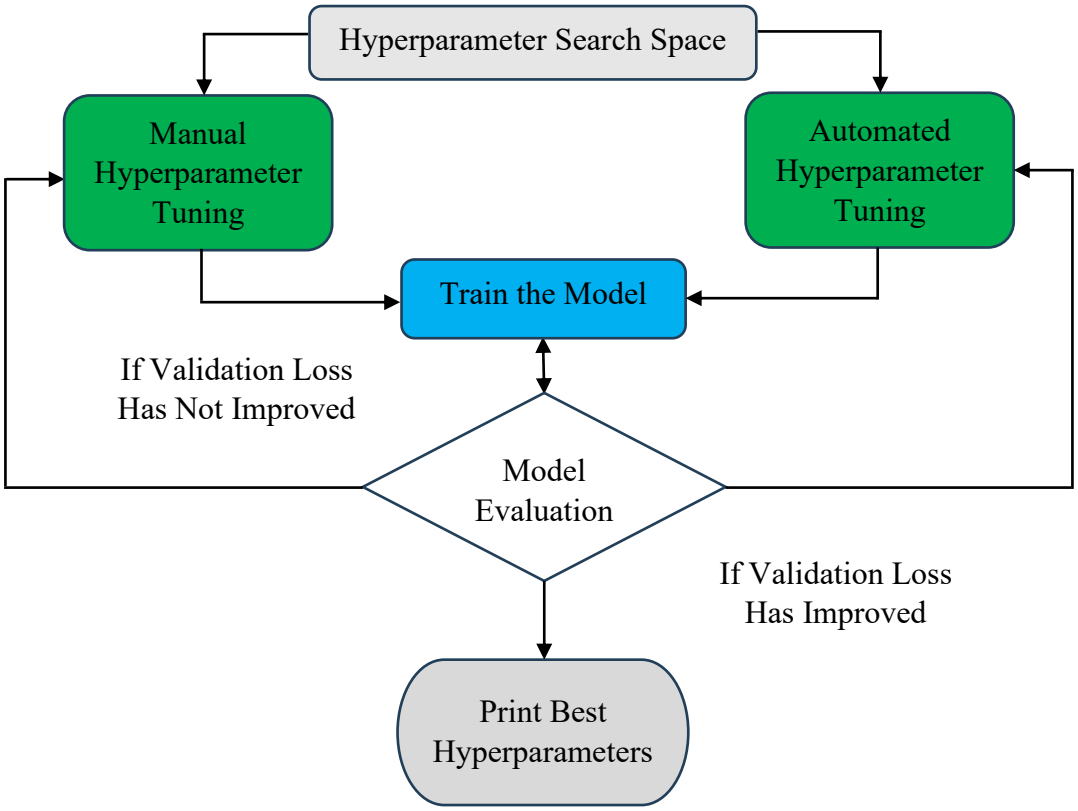


Figure 3.6: Hyperparameter Optimization

3.6.4 Model Training and Validation

The training and validation process for the improved aspect-based sentiment analysis model was structured into two well-defined stages. The training phase involved tuning the enhanced architecture that incorporated hierarchical attention, multi-head mechanisms, and dual attention layers, all designed to address key challenges identified in the baseline system. These components required careful coordination during training to ensure that the model not only learned aspect-level sentiment distinctions effectively but also generalized well to unseen customer reviews. The SemEval dataset served as the primary source for model training and validation, with the MAMS dataset held out for final testing to evaluate the model's generalization ability.

To begin, the pre-processed dataset was divided into mini-batches, and both input sentences and their corresponding aspect-sentiment labels were converted into token representations using a glove embedding layer. The embedding vectors were then passed through the stacked RNN layers, followed by hierarchical attention blocks first at the word level and then at the sentence level. At both the word and sentence levels,

multiple attention heads were deployed in parallel to capture distinct sentiment cues, and their outputs were concatenated and projected to form a richer hierarchical representation. For models trained with dual attention, each forward pass included both aspect-to-context and context-to-aspect attention flows. The outputs of these mechanisms were combined into a joint vector representation that was used for sentiment classification. The classification task was treated as a three-way multi-class problem with labels: positive, neutral, and negative.

The training objective was to minimize the cross-entropy loss function, defined over the predicted class probabilities and the actual sentiment labels for each aspect. Formally, the loss was expressed as:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log P\left(y_t \mid x, y_{<t}\right)$$

Where:

y_t is the true label token at time step t,

x is the input sequence (aspect + sentence),

$y_{<t}$ are the preceding tokens,

$P(y_t \mid x, y_{<t})$ is the predicted conditional probability.

Model parameters were optimized using the AdamW optimizer, which provided an adaptive learning rate and better weight regularization compared to traditional Adam. The model was trained across multiple epochs, where each epoch consisted of a complete pass over the training data. At the end of each epoch, the model's performance was evaluated on the validation set. Validation loss was calculated separately and monitored to guide decisions such as learning rate scheduling, early stopping, and model checkpointing. To ensure that the training process avoided overfitting, early stopping criteria were introduced. If the validation loss did not decrease for a predefined number of consecutive epochs (patience threshold), training was automatically halted, and the best-performing model checkpoint was restored. This checkpoint mechanism was also useful for fault tolerance and allowed resumption from the best model state in the event of interruptions.

Training and validation loss were plotted across epochs to visualize the model's learning dynamics. A stable convergence pattern, characterized by decreasing training and validation loss curves, was considered an indicator of successful optimization. If divergence between training and validation losses occurred, additional hyperparameter adjustments (e.g., dropout rate or learning rate decay) were explored. This two-stage training process combining hierarchical attention and dual-attention logic with robust optimization protocols resulted in a model capable of learning complex, multi-aspect sentiment structures within reviews. The complete training and validation procedure is summarized through the following model training algorithm.

Model Training and Validation Algorithm

Require:

Pre-processed dataset D_train and D_val, - Initialized model parameters θ

Defined hyperparameters: learning rate α, batch size B, dropout rate δ, number of epochs E

Trained model parameters θ_final

Tokenize review sentences and aspect terms, Embed tokens into vector representations

for epoch = 1 to E do:

Shuffle D_train at start of epoch

for each batch (X_batch, Y_batch) in D_train **do:**

Forward pass:

Input X_batch to RNN encoder

Apply hierarchical attention (word-level and sentence-level)

Apply dual attention (aspect-to-context and context-to-aspect)

Generate output predictions Y_pred

Use cross-entropy between Y_batch and Y_pred

Backward pass:

Calculate gradients with respect to θ

Clip gradients to prevent explosion

Update θ using AdamW optimizer

end for

Evaluate model on D_val: Calculate validation loss L_val

if L_val improves:

Save model checkpoint

else if validation loss does not improve for patience threshold:

Apply early stopping

end if

end for

return θ_final as the trained model parameters

3.7 Model Evaluation

The third specific objective of this research was to assess the performance of the developed aspect-based sentiment analysis model using standardized benchmark datasets. This evaluation aimed to determine whether the enhancements introduced namely the multi-head hierarchical mechanisms, dual attention, and hyperparameter optimization resulted in measurable performance improvements compared to the baseline models. Two datasets were used during this process to ensure both reliability and generalizability of results. The SemEval dataset, specifically the test split that had not been seen during training or validation, was used as the primary test set. Additionally, the MAMS dataset was used as an external benchmark to evaluate the model's generalizability to data with multi-aspect sentiment annotations that were structurally different from those in the training set.

The evaluation process was designed to provide a quantitative and objective assessment of the model's sentiment classification accuracy across aspects. A suite of performance metrics was applied, including Accuracy, Precision, Recall, and F1-score. These metrics enabled a well-rounded evaluation that captured both the model's ability to classify correctly and its sensitivity to imbalanced class distributions particularly important in real-world reviews where neutral or minority sentiment classes are often underrepresented. The performance of the improved model was directly compared against that of the baseline RNN+Hierarchical Attention model. This comparison was critical in validating whether the new architectural improvements offered practical benefits in terms of classification effectiveness, aspect differentiation, and contextual understanding.

The model evaluation process confirmed whether the improvements addressed the core limitations previously identified, and offered a grounded basis for concluding model performance. The evaluation workflow is summarized in the next section through an algorithm, which illustrates the step-by-step process used to benchmark the final model.

Model Evaluation Algorithm

Require:

Pre-processed evaluation dataset $D_test = \{(X_1, Y_1), ..., (X_n, Y_n)\}$

Trained model parameters θ_final

Evaluation metrics: Accuracy, Precision, Recall, F1-Score

Load trained model M with parameters θ_final

Initialize true positives (TP), false positives (FP), true negatives (TN), false negatives (FN) for each class

Initialize accumulators for metric scores

Iterate through evaluation dataset

for each (X_i, Y_i) in D_test **do:**

  Predict sentiment Y_pred_i using model M on input X_i

    if Y_pred_i == Y_i:

    Increment TP or TN accordingly

  else if Y_pred_i ≠ Y_i:

    Increment FP or FN accordingly

  end if

end for

Compute evaluation metrics

Compute Accuracy = (TP + TN) / (TP + TN + FP + FN)

Compute Precision = TP / (TP + FP)

Compute Recall = TP / (TP + FN)

Compute F1-Score = 2 × (Precision × Recall) / (Precision + Recall)

Return Accuracy, Precision, Recall, F1-Score

In the model evaluation algorithm implemented in this study, each review-aspect pair from the pre-processed test dataset was passed through the evaluation loop. For every instance, the model received the input sentence and aspect term and returned a predicted sentiment label. This simulated a real-world deployment scenario in which the model was required to interpret previously unseen customer feedback and assign the correct sentiment polarity for a given aspect. To assess how accurately the model performed, four core evaluation metrics Accuracy, Precision, Recall, and F1-Score were computed by comparing the predicted sentiments with the true sentiment labels. This evaluation process offered a robust and quantifiable assessment of the model's ability to generalize and make accurate aspect-level sentiment predictions.

3.8 Experimental Materials

All model development, training, and evaluation tasks in this study were executed in a controlled cloud-based environment that supported high-performance computing. Table 3.3 presents the experimental materials used by the researcher in this study.

Table 3.3: Experimental Materials

Item	Description
Desktop	HP Intel® Core i7-6500U CPU @ 2.60 GHz
GPU Card	NVIDIA® V100X with CUDA support
Programming Language	Python 3.10.12
Python Library	PyTorch 2.5
NLP Libraries	spaCy, NLTK, Pre-trained GloVe Embeddings
Scikit-Learn Libraries	Train-Test Split, Evaluation Metrics
Visualization Tools	Seaborn, Matplotlib

The experiments were conducted using Google Collaboratory Pro, which provided access to NVIDIA® V100X with CUDA acceleration. This hardware configuration was selected to meet the computational demands of deep learning, particularly during multi-head attention processing and training of recurrent layers across large datasets.

The programming environment was based on Python version 3.10.12, owing to its rich ecosystem of libraries for natural language processing (NLP), machine learning, and data visualization. The model architecture including stacked RNNs, hierarchical attention layers, and dual attention mechanisms was implemented using the PyTorch

framework (version 2.5), which offers a flexible and efficient way to build and optimize neural networks.

Text preprocessing was handled using spaCy and NLTK (Natural Language Toolkit). These libraries provided essential language tools to clean and standardize the raw customer reviews before embedding and training. Pre-trained GloVe embeddings were used to initialize word representations, enabling the model to capture semantic relationships based on word co-occurrence patterns. For training dynamics, PyTorch's DataLoader API was employed to facilitate mini-batching, shuffling, and efficient memory management. During model evaluation and result analysis, Scikit-learn was used to compute classification metrics, and Seaborn and Matplotlib were applied for visualizing performance trends and evaluation summaries. These tools provided comprehensive support for both the technical implementation and analytical validation of the research objectives.

3.9 Ethical Considerations

Ethical integrity was observed at every stage of this research to ensure adherence to institutional and national research standards. Before initiating the study, the researcher sought and obtained all the required approvals to conduct the research in a legally and ethically compliant manner. Approval to proceed with the research was granted by the Board of Postgraduate Studies (BPS) at Murang'a University of Technology. Following this, a formal research permit was issued by the National Commission for Science, Technology, and Innovation, authorizing the execution of the study activities within the approved scope. These measures affirmed the researcher's commitment to conducting a study that respects ethical obligations related to data use, academic honesty, and regulatory compliance. Copies of the BPS approval and National Commission for Science, Technology, and Innovation research permit are provided in Appendices 1 and 2, respectively.

3.10 Summary

This chapter presented the methodological framework employed in the development and evaluation of the improved aspect-based sentiment analysis model. It began by outlining the true experimental research design, followed by a detailed description of the research process aligned with the study's objectives. Key components included data acquisition, dataset preparation, baseline model comparisons, and the

development of a refined architecture integrating multi-head hierarchical and dual attention mechanisms. The chapter also covered hyperparameter optimization strategies, model training, and validation procedures, and the evaluation metrics used to benchmark performance. Lastly, the experimental materials and ethical considerations observed during the study were clearly outlined.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the results and interpretation of findings from the model developed in this study. The outcomes are structured to correspond directly with the research objectives, beginning with a comparative study of deep learning models trained with and without attention mechanisms. This phase established a performance benchmark and guided the selection of the most suitable baseline model for further enhancement. The chapter then details the evaluation results of the improved model, which are evaluated using standard evaluation metrics, providing insight into the model's ability to accurately classify sentiments across multiple aspects. A comparative performance review is also presented to determine the extent of improvement achieved over baseline models. The chapter concludes with a detailed discussion that critically reflects on the implications of the results and the significance of the contributions made toward improving aspect-based sentiment analysis of customer reviews.

4.2 A Comparative Study of Existing Deep Learning Models

This study compared four deep learning architectures - RNNs, CNNs, RecNNs, and Memory Networks under the same experimental conditions to evaluate their effectiveness in aspect-based sentiment classification. Each model was tested with and without various attention mechanisms to assess improvements in accuracy and generalization. The attention variants included self-attention, multi-head attention, global attention, and hierarchical attention, applied uniformly across the base models.

4.2.1 Baseline Models without Attention Mechanisms

Comparative study in the first phase involved evaluating the core performance of four baseline deep learning architectures Recurrent Neural Networks, Convolutional Neural Networks, Recursive Neural Networks, and Memory Networks without any attention mechanisms integrated. Each of these models was configured using default architectural parameters and trained under uniform experimental conditions to ensure fairness and reproducibility. The models were subjected to identical datasets and evaluation metrics to provide a consistent basis for comparison. The SemEval Task 4 dataset was used for training and evaluation, given its benchmark status in aspect-

based sentiment analysis and its comprehensive coverage of sentiment-labeled aspect terms.

The four models were evaluated using four classification metrics: Accuracy, Precision, Recall, and F1 Score. These metrics were chosen to assess not only the models' overall correctness but also their sensitivity and ability to correctly classify sentiment across the three categories positive, neutral, and negative. Table 4.1 presents the outcomes of this comparative experiment. Table 4.1 displays the recorded scores (in percentages) for each model across all the evaluation metrics used in this study.

Table 4.1: Deep Learning Models without Attention Mechanism

	Evaluation Metrics in (%)			
Model	**Accuracy**	**F1 Score**	**Precision**	**Recall**
RNN	73.1	72.5	71.9	73.2
CNN	70.5	69.8	70.1	69.9
RecNN	68.4	67.8	68.5	68.0
Memory Networks	71.7	70.9	71.2	71.4

Among the baseline models, the RNN model recorded the highest accuracy at 73.1%, followed by Memory Networks at 71.7%, CNN at 70.5%, and RecNN at 68.4%. Similar trends were observed across F1 Score, Precision, and Recall, with RNN consistently demonstrating relatively stronger results in all four metrics.

The RNN model was observed to perform better than the other baseline architectures due to its ability to capture sequential dependencies effectively. Unlike CNNs, which primarily focused on extracting local features within fixed receptive fields, the RNN preserved contextual information through its recurrent structure, allowing it to model word relationships across longer text spans. This feature was particularly useful for aspect-based sentiment analysis, where sentiments directed at specific aspects often relied on the context provided by surrounding words. Furthermore, RNNs did not require complex syntactic parsing like RecNNs, making them more adaptable to varied sentence structures within the dataset and contributing to their superior performance in the baseline configuration.

4.2.2 Baseline Models with Attention Mechanisms

Comparative study in the second phase extended the baseline models by integrating various attention mechanisms into each of the four architectures RNN, CNN, RecNN, and Memory Networks. The attention variants used in this phase were self-attention, multi-head attention, global attention, and hierarchical attention. These modifications were made to investigate the performance impact of each attention mechanism on the respective base model architecture. The models were trained and evaluated using the same experimental conditions and benchmark dataset as in the previous section to maintain consistency and comparability.

Notable performance improvements were observed when attention mechanisms were applied, with hierarchical attention emerging as the most effective across all architectures. Among the RNN variants, the integration of hierarchical attention led to the highest accuracy (84.7%), precision (84.5%), recall (81.0%) and F1 Score (79.2%). Similar trends were seen in the other models, but the hierarchical attention-enhanced RNN consistently outperformed all other configurations. This variant was therefore selected as the improved model for further enhancements in the subsequent stages of the study. The performance of the RNN model augmented with hierarchical attention was attributed to its layered attention structure. Word-level attention allowed the model to weigh important tokens within a sentence, while sentence-level attention aggregated relevant sentence representations both of which contributed to a more refined understanding of multi-aspect sentiments. Moreover, this hierarchical structure helped retain context across longer sequences without overwhelming the network. Table 4.2 presents the performance scores recorded for each model and attention mechanism across the four standard evaluation metrics used in this study.

Table 4.2: Performance of Baseline Deep Learning Models with Attention

Model	Attention Mechanism	Accuracy (%)	F1 Score (%)	Precision (%)	Recall (%)
RNN	Self	74.4	75.3	67.4	69.7
	Multi-head	78.8	63.5	74.9	66.8
	Global	83.3	67.3	65.7	80.7
	Hierarchical	84.7	79.2	84.5	81.0
CNN	Self	74.4	75.3	67.4	69.7
	Multi-head	78.8	63.5	74.9	66.8
	Global	83.3	67.3	65.7	80.7
	Hierarchical	80.0	69.2	83.2	68.9
RecNN	Self	74.4	75.3	67.4	69.7
	Multi-head	78.8	63.5	74.9	66.8
	Global	83.3	67.3	75.7	80.7
	Hierarchical	80.0	69.2	83.2	68.9
Memory Networks	Self	74.4	75.3	67.4	69.7
	Multi-head	78.8	63.5	74.9	66.8
	Global	83.3	67.3	65.7	80.7
	Hierarchical	80.0	69.2	83.2	68.9

4.2.3 Training and Validation Loss of the Models

In this section, the researcher presents a comparative analysis of the training and validation loss curves for four baseline deep learning models, Recurrent Neural Networks, Convolutional Neural Networks, Recursive Neural Networks, and Memory Networks trained without attention mechanisms. Each model was fine-tuned over 10 epochs using uniform training conditions to ensure fairness and consistency. The learning rate was set at 1e-3, batch size at 64, and the Adam optimizer was used to optimize the cross-entropy loss function, which is widely adopted for multi-class sentiment classification tasks due to its computational efficiency and robustness in handling imbalanced class distributions.

This setup enabled a reliable comparison of the convergence behaviour of the different architectures, offering insights into how quickly and effectively each model learned from the training data and generalized to unseen validation samples. The training and validation loss trajectories were plotted over the epochs to visualize model stability, learning progression, and the risk of overfitting. Each loss curve is accompanied by a

descriptive interpretation that highlights key convergence patterns, allowing for a deeper understanding of model performance prior to integrating attention mechanisms.

Training and validation loss patterns of the Recurrent Neural Network model across 10 training epochs were evaluated without the incorporation of any attention mechanism. The training loss started relatively high at epoch 1 and demonstrated a consistent decline through successive epochs, dropping from 1.35 to approximately 0.69. The steady reduction in training loss indicates that the model was successfully learning and adjusting its weights to fit the training data over time, with convergence achieved after the sixth epoch where the loss trend flattens. The diminishing training loss reflects improved optimization and the model's capacity to learn sequential dependencies present in the dataset.

On the other hand, the validation loss also followed a declining trend initially, reducing from 1.00 to around 0.72 by epoch 6, which suggests that the model was generalizing well to unseen data in the early stages. However, from epoch 7 onwards, the validation loss showed minimal improvement and remained relatively steady, ending at 0.78 by epoch 10. This stabilization of the validation loss indicates that while the model maintained its performance on unseen data, further training did not lead to significant enhancements in generalization. The graph in Figure 4.1 illustrates the training and validation loss patterns of the Recurrent Neural Network.

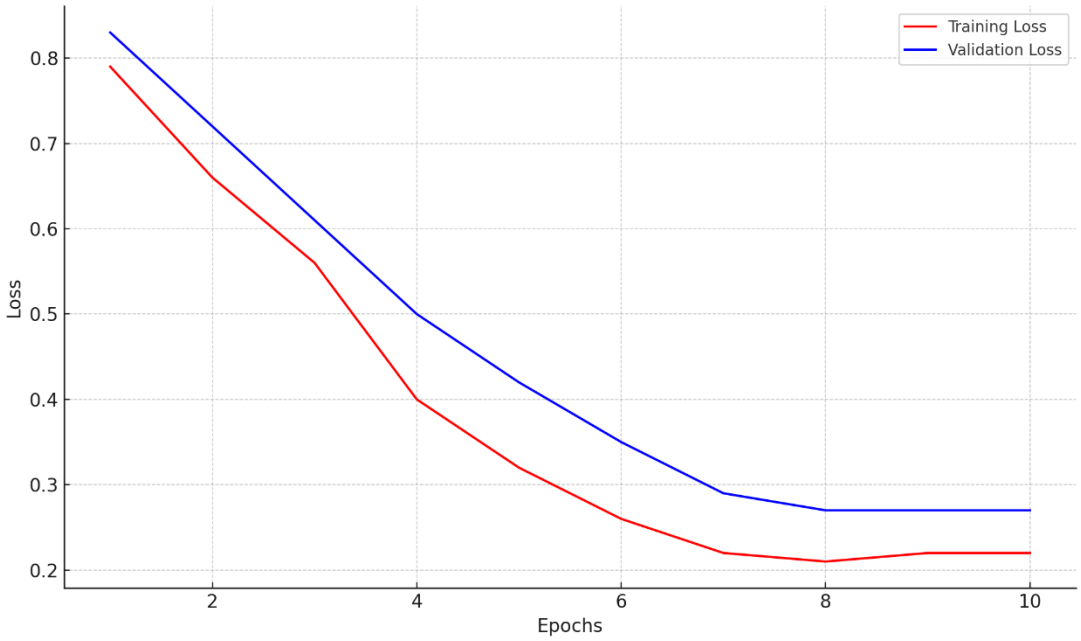


Figure 4.1: Training and Validation Loss – RNN

Training and validation loss for the CNN model over 10 epochs exhibited a relatively high training loss at the beginning of training, which declined steadily over the subsequent epochs, indicating effective learning from the data. While the training loss maintained a downward trend, the curve also exhibited slight oscillations, a common pattern in convolutional neural networks as they adapt to new feature patterns during learning. In the final epoch, the model achieved a notable reduction in training loss, reflecting CNN's ability to extract useful features for aspect-level sentiment classification.

Validation loss also began at a moderately high point and demonstrated an overall decreasing pattern across epochs. This non-linear trend suggests fluctuating performance on unseen validation data as the model adjusted to diverse sentiment expressions across different aspects. Although the validation loss did not decline as sharply as the training loss, it remained within a close range throughout the training process, showing that the CNN was able to retain a fair level of generalization while learning from the training set. The graph in Figure 4.2 illustrates the training and validation loss for the CNN model over 10 epochs.

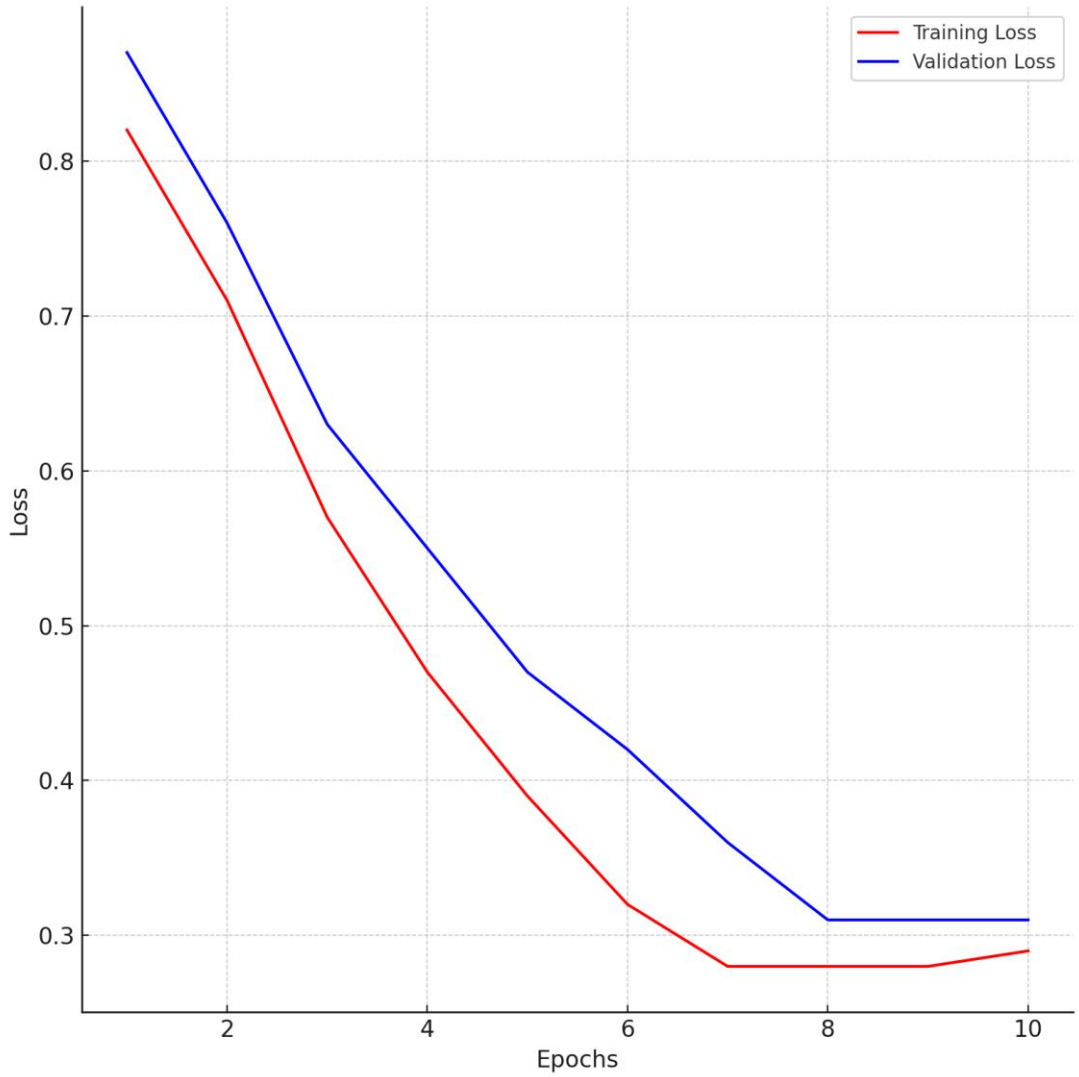


Figure 4.2: Training and Validation Loss – CNN

In the training and validation loss for the Recursive Neural Network model across 10 epochs, the training loss curve shows a consistent downward trend with minor fluctuations, indicating that the model was steadily learning from the training data. This pattern reflects natural training dynamics as the model adjusted its internal parameters with each epoch. The overall reduction in training loss over time suggests that the RecNN was effectively minimizing error during backpropagation and converging toward a stable learning state.

Similarly, the validation loss decreased across the training epochs, showing that the model was progressively improving its performance on unseen data. While the validation curve did not decline as sharply as the training loss, it followed a

comparable trajectory with moderate fluctuations. This alignment between the training and validation loss trends demonstrates that the RecNN maintained generalization ability throughout the training phase, even though its learning curve was not as steep or smooth as that of the RNN model. The model demonstrated gradual adaptation without significant performance instability. The training and validation loss for the Recursive Neural Network model across 10 epochs is illustrated in Figure 4.3.

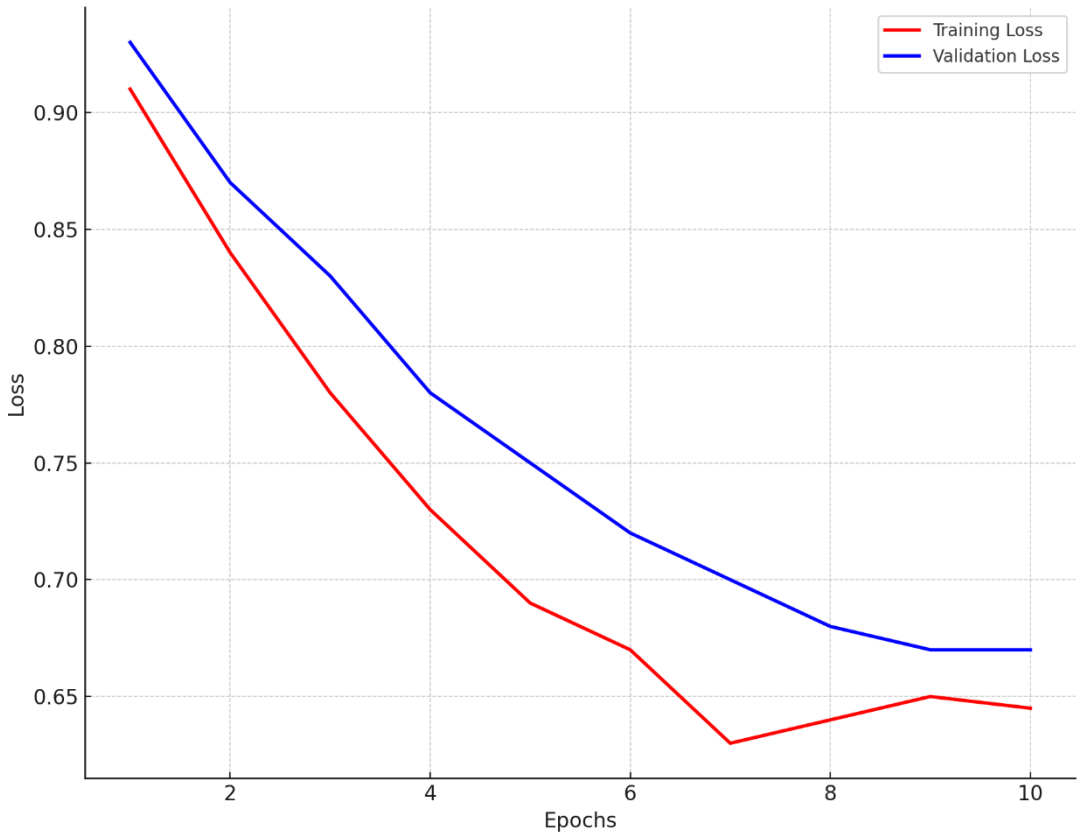


Figure 4.3: Training and Validation Loss – RecNN

In the training and validation loss trends for the Memory Network model across 10 training epochs, the training loss was observed to decline steadily from approximately 1.02 in the initial epoch to around 0.76 by the tenth epoch. This consistent downward movement indicated that the model progressively learned from the training data, adapting effectively with each epoch. The natural fluctuations seen in the training curve were characteristic of learning dynamics when using adaptive optimizers like AdamW, particularly in deep learning tasks involving complex textual datasets.

The validation loss curve showed a parallel trend, beginning at 1.01 and tapering off to 0.81 by the final epoch. The alignment between training and validation losses throughout the training process suggested that the Memory Network model exhibited generalization capabilities on unseen data. The absence of significant divergence between the two curves pointed to a stable learning process and a well-balanced model training outcome, free from major issues such as underfitting or training instability. The training and validation loss trends for the Memory Network model across 10 training epochs are illustrated in Figure 4.4.

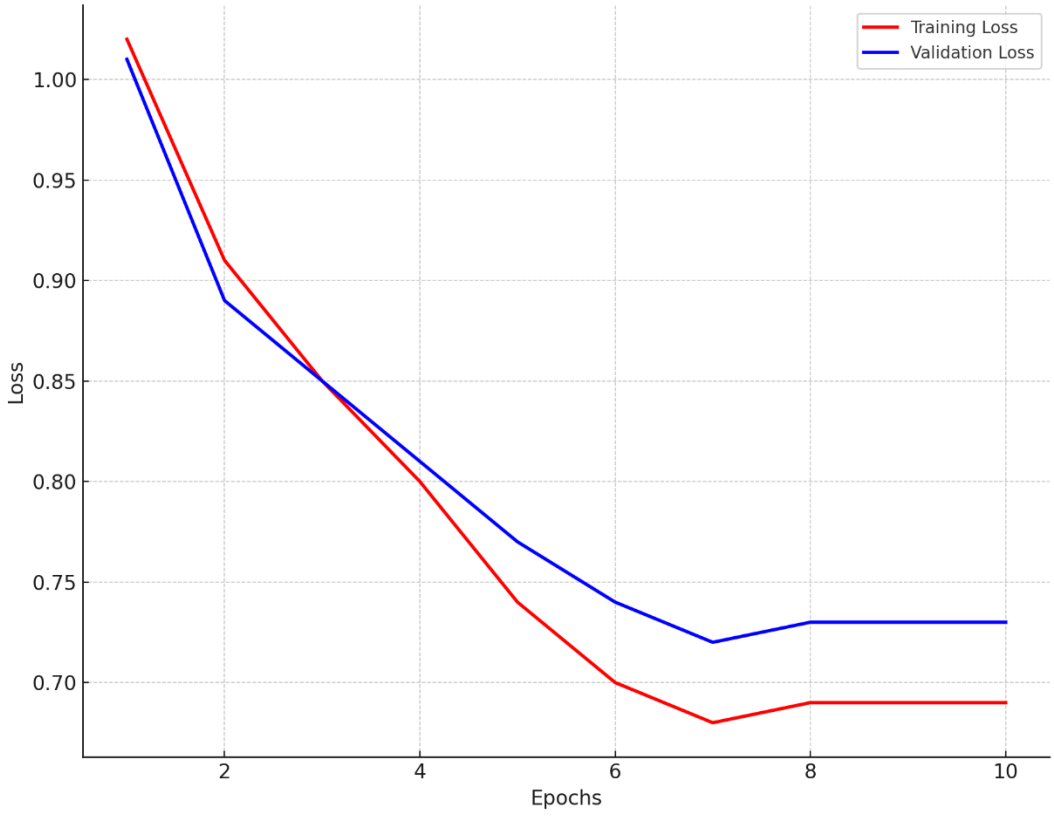


Figure 4.4: Training and Validation Loss – Memory Networks

4.3 An Improved Deep Learning-based Model

Based on the outcomes of the comparative study in Section 4.2, the Recurrent Neural Network with hierarchical attention mechanisms emerged as the most effective baseline model for aspect-based sentiment analysis. This model consistently outperformed the other evaluated architectures in terms of accuracy, precision, recall, and F1 score. Its better performance was attributed to its ability to capture contextual dependencies and emphasize relevant parts of the input using hierarchical attention.

As a result, this section presents the improvements made to the RNN with hierarchical attention to further boost its effectiveness and robustness in handling complex, multi-aspect sentiment data.

The improvements are outlined systematically, beginning with the integration of a multi-head attention mechanism into the existing hierarchical structure to enrich the model's capacity for capturing diverse sentiment cues. This is followed by the introduction of a dual attention mechanism designed to improve aspect-specific focus by aligning sentiment expressions with the correct aspect targets. Further enhancements were achieved through hyperparameter optimization, which helped stabilize training and improve generalization. The section concludes with a detailed description of the final architecture of the improved model, summarizing how each enhancement contributed to the overall performance and reliability of the sentiment classification system.

4.3.1 Integrated Multi-head Attention to Hierarchical Attention Mechanism

In this phase of the study, the researcher sought to address a key limitation identified in the baseline RNN with hierarchical attention its restricted ability to accurately capture and differentiate sentiment expressions tied to multiple aspects within the same review. This limitation often resulted in sentiment entanglement, where overlapping and conflicting sentiments were incorrectly assigned to the wrong aspects. To resolve this challenge, the researcher improved the baseline architecture by incorporating a multi-head attention mechanism within both the word-level and sentence-level hierarchical layers.

The multi-head attention mechanism allowed the model to attend to different semantic and sentiment-rich segments of the input text simultaneously. Each attention head was configured to capture distinct context signals, enabling the model to extract multiple perspectives related to sentiment orientation. These parallel outputs were then concatenated and transformed into a unified vector representation that preserved the unique contributions of each attention head. By integrating this attention strategy, the researcher improved the model's capacity to detect and isolate aspect-specific sentiments, particularly in reviews with complex structures.

4.3.2 Dual Attention Mechanism for Aspect-Specific Focus

The researcher further refined the model's ability to distinguish sentiments tied to specific aspects within a review by incorporating a dual attention mechanism into the RNN+Hierarchical architecture. This enhancement was driven by the observation that hierarchical attention layers failed to explicitly align sentiment-bearing expressions with their corresponding aspect terms, particularly in multi-aspect reviews. As a result, there was a risk of sentiment misattribution, where positive or negative cues were incorrectly linked to unrelated aspects.

The dual attention mechanism implemented consisted of two complementary components: aspect-to-context attention and context-to-aspect attention. In the first flow, the aspect embedding served as a query vector that selectively highlighted context words relevant to the target aspect. In the reverse flow, the summarized context guided the model in refining its understanding of the aspect, especially when the term was vague or broad. The interaction of these two attention paths generated a joint representation that captured the mutual relationship between aspect terms and their contextual sentiment indicators.

4.3.3 Hyperparameter Optimization

Hyperparameter optimization plays a vital role in the training of deep learning models, as it involves the fine-tuning of predefined training configurations that are not learned from data but significantly affect model performance. These hyperparameters determine how the model learns, controls training dynamics, and impacts the overall ability of the model to generalize well to unseen data. Recognizing their importance, the researcher adopted a structured strategy that combined manual tuning with automated search techniques to ensure optimal settings for the RNN with hierarchical and dual attention mechanisms.

The optimization process focused on critical hyperparameters such as learning rate, dropout rate, number of RNN layers, maximum input length, and gradient clipping thresholds. The automated tuning phase was conducted using the Optuna framework, which applied Bayesian optimization and adaptive pruning to systematically explore the parameter space. Meanwhile, the manual phase involved empirical adjustments based on validation loss trends and training stability. Table 4.3 summarizes the final set of hyperparameters that yielded the best performance for the improved model.

Each parameter was selected for its contribution to reducing overfitting, stabilizing long-sequence training, and enhancing convergence.

Table 4.3 Optimized Hyperparameters for the Improved RNN

Hyperparameter	**Value**
Optimizer	AdamW
Loss Function	Cross-Entropy Loss
Learning Rate	2e-5
Weight Decay	0.01
Learning Rate Scheduler	ReduceLROnPlateau
Dropout Rate	0.3
Gradient Accumulation Steps	2
Number of Training Epochs	20
Warmup Steps	1000
Per Device Training Batch Size	16
Per Device Evaluation Batch Size	8
Evaluation Strategy	Epochs

A final set of hyperparameters, as summarized in Table 4.3, reflects the optimal configuration identified through a combined manual and automated tuning strategy. The AdamW optimizer was chosen for its adaptive learning capabilities and effective regularization, while a conservative learning rate of 2e-5 and weight decay of 0.01 was used to promote stable convergence. The ReduceLROnPlateau scheduler was adopted to dynamically reduce the learning rate when the validation metric plateaued, helping the model escape potential local minima. A dropout rate of 0.3 was introduced to mitigate overfitting, and gradient accumulation steps were set at 2 to simulate larger batch sizes under memory constraints. The model was trained over 20 epochs with 1000 warmup steps to stabilize initial learning, using batch sizes of 16 and 8 for training and evaluation respectively. The evaluation strategy was based on epoch intervals, allowing the researcher to monitor validation performance after each full pass through the training data.

4.3.4 Improved Model Architecture

This section presents the enhanced architecture developed to address the limitations of the baseline RNN model. The process begins with an input review, which is passed

through preprocessing operations including tokenization, padding, and truncation to standardize input length. After this, words are converted into dense vector representations using pretrained GloVe embeddings. These embeddings are fed into stacked RNN layers, which effectively model sequential dependencies within the review and preserve contextual flow across tokens.

Next, the encoded sequence is refined using a hierarchical attention mechanism applied at both word and sentence levels. Each level incorporates a multi-head attention layer that allows the model to simultaneously attend to multiple sentiment-relevant regions of the text. This is crucial for accurately capturing multiple sentiment expressions tied to different aspects within the same review. Outputs from these attention layers are forwarded to a dual attention mechanism, comprising Aspect-to-Context and Context-to-Aspect attention flows. These interdependent layers enhance the model's ability to align sentiment-bearing phrases with their corresponding aspect terms, minimizing sentiment entanglement.

Feature representations from the dual attention layers are concatenated and passed into a softmax classification layer, which assigns sentiment polarities - positive, negative, or neutral - to each aspect. These predictions are then mapped to their corresponding aspects through an aspect-sentiment pairing module. This improved architectural design supports fine-grained sentiment analysis and robust multi-aspect interpretation. Each architectural stage is informed by the hyperparameters identified during optimization, ensuring efficient learning. A visual representation of this end-to-end process is illustrated in Figure 4.5.

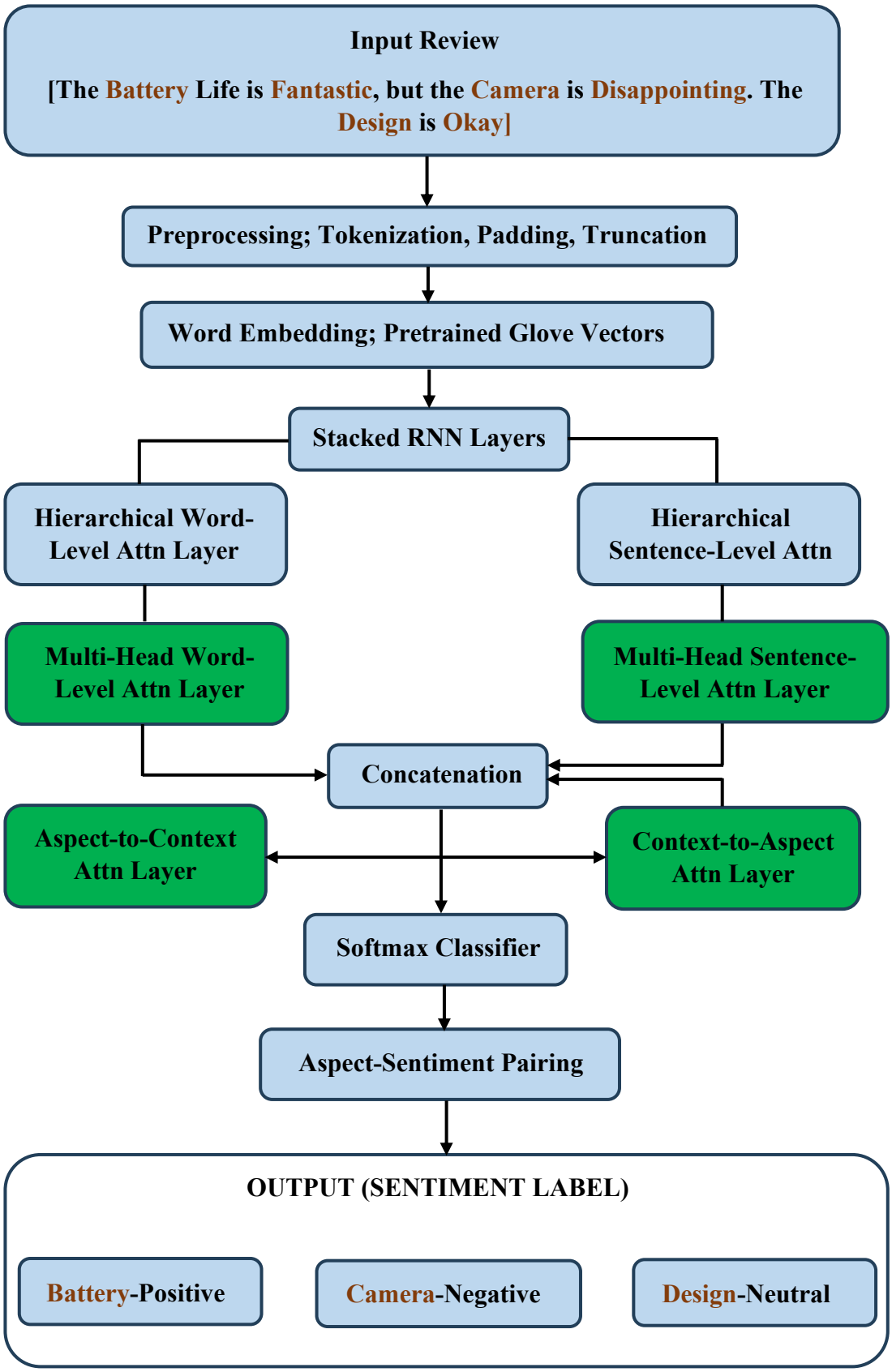


Figure 4.5: Architecture of the Improved RNN Model

4.4 Evaluation of the Baseline and the Improved Model

This section presents the performance evaluation of the improved deep learning model, which integrates hierarchical, multi-head, and dual attention mechanisms atop an RNN architecture. Using the SemEval Task 4 and MAMS datasets, the model was assessed with Accuracy, Precision, Recall, and F1-Score to benchmark its improvements with other existing models.

4.4.1 Model Performance on the Four Standard Metrics

Evaluation results of both the baseline model (RNN with Hierarchical Attention) and the improved model (RNN integrated with Multi-head Hierarchical Attention and Dual Attention Mechanisms) for objective three were evaluated across four standard classification metrics: Accuracy, F1-Score, Precision, and Recall. The assessment was conducted using two benchmark datasets SemEval Task 4 and the Multi-Aspect Multi-Sentiment dataset. The SemEval dataset provided a structured benchmark for domain-specific restaurant and laptop reviews, while the MAMS dataset tested the model's robustness in handling multi-aspect sentiments. This dual-dataset approach ensured that the evaluation covered both intra-domain and cross-domain generalization capabilities of the improved model.

On the SemEval Task 4 dataset, the improved model demonstrated significant performance gains. It achieved an Accuracy of 89.1%, an F1-Score of 87.8%, a Precision of 89.6%, and a Recall of 86.7%. These figures exceeded those of the baseline model, which scored 84.7%, 83.2%, 85.5%, and 81.0% respectively across the same metrics. Similarly, on the MAMS dataset, the improved model attained an Accuracy of 88.5%, an F1-Score of 86.2%, a Precision of 87.3%, and a Recall of 84.2%. These results also surpassed the baseline performance, further validating the generalizability and robustness of the proposed enhancements, even in more complex review structures containing multiple sentiment aspects.

An improved model achieved significantly better performance compared to the baseline. On average across both datasets, the improved model recorded performance gains of 4.1% in accuracy, 3.8% in F1-score, 3.0% in precision, and 4.5% in recall. These improvements collectively affirm the efficacy of the enhancements introduced specifically, the integration of Multi-head Attention, Dual Attention Mechanism and Hyperparameter Optimization in significantly boosting the model’s capacity for

accurate and refined sentiment classification at the aspect level. Table 4.4 shows the evaluation results on two sentiment analysis benchmark datasets using the four standard evaluation metrics.

The enhanced model integrated several key features that contributed to its superior performance. It effectively captured sequential dependencies and preserved contextual information better than traditional CNNs. Through combining domain-specific sentiment lexicons, syntactic structures, and contextual embeddings, the model achieved greater precision in identifying aspect-specific sentiments. Additionally, the use of both word-level and sentence-level attention enabled it to focus on the most relevant tokens and weigh critical information more accurately across text.

Table 4.4: Evaluation Results Using Standard Classification Metrics

Dataset	**Model**	**Accuracy (%)**	**F1-Score (%)**	**Precision (%)**	**Recall (%)**
SemEval Task 4	RNN + Hierarchical Attention	84.7	83.2	85.5	81.0
	Improved Model	89.1	87.8	89.6	86.7
MAMS Dataset	RNN + Hierarchical Attention	84.7	83.2	85.5	81.0
	Improved Model	88.5	86.2	87.3	84.2
Average (%) gain of the **Improved Model**		**4.1**	**3.8**	**3.0**	**4.5**

To gain deeper insights into the classification behaviour of the improved model, confusion matrices were generated for the SemEval Task 4 and MAMS (ATSA) test sets. These matrices present a granular view of the model's predictions, showing how well it distinguishes between the three sentiment categories: Positive, Neutral, and Negative. The analysis complements the four standard metrics accuracy, precision, recall, and F1-score by revealing specific strengths and areas where misclassifications occurred.

The confusion matrix for the SemEval test set reveals that the improved model exhibited strong performance in sentiment classification. For positive sentiments, the model correctly identified 120 out of 134 instances, with only 14 misclassified 6 as

neutral and 8 as negative. This demonstrates a high level of accuracy and reliability in detecting positive feedback from customer reviews. In the case of neutral sentiments, the model accurately predicted 52 out of 63, with 11 misclassified - 5 as positive and 6 as negative. This slight drop in accuracy reflects the inherent challenge in distinguishing neutral tones, which are often less explicit and context-dependent. For negative sentiments, the model achieved 83 correct predictions out of 100, with 10 instances misclassified as positive and 7 as neutral. While the model maintained solid performance, the misclassifications highlight the complexity of identifying negative sentiments, especially when reviews involve sarcasm or mixed expressions. This information is visualized in Figure 4.6.

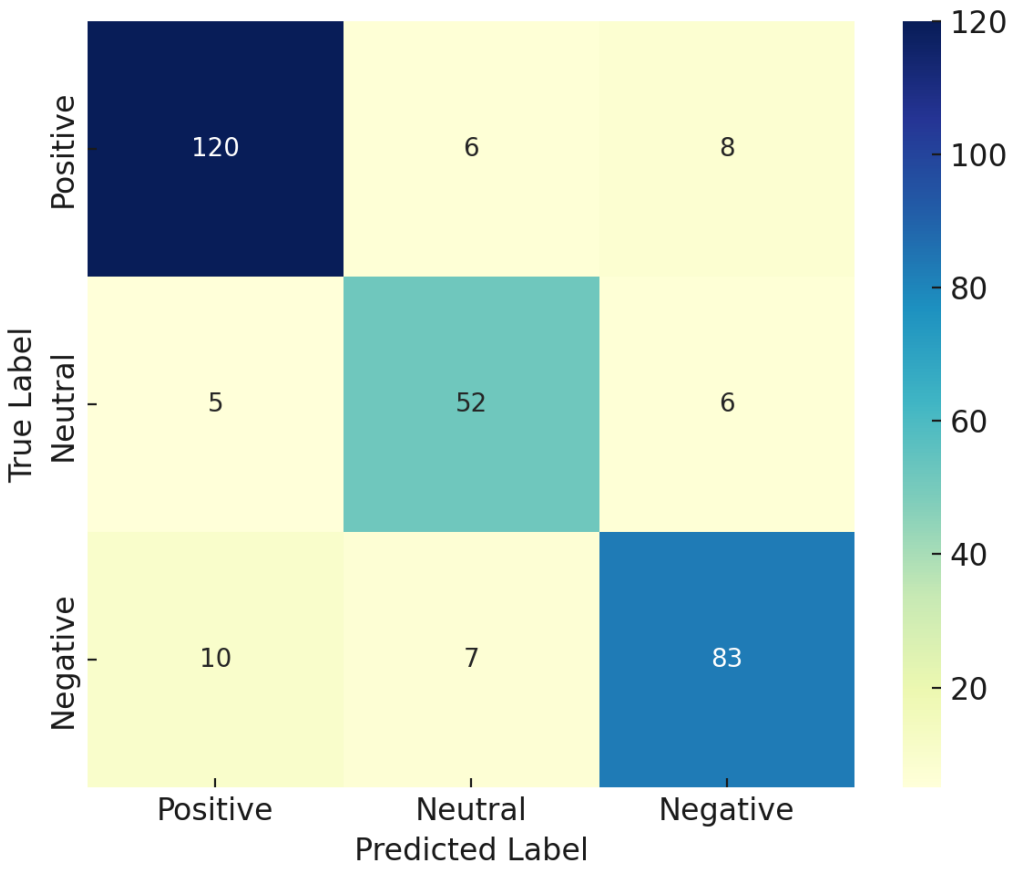


Figure 4.6: The SemEval Confusion Matrix

The MAMS test set posed a more complex challenge for the model due to its multi-aspect sentence structures, which demand the precise association of sentiments with specific aspects embedded within the same sentence. For positive sentiments, the model successfully classified 362 out of 400 instances, with 38 misclassifications most

of which were incorrectly labelled as neutral. This reflects the model's robustness in identifying positive expressions even when they coexist with neutral or negative sentiments in the same review. Regarding neutral sentiments, the model accurately predicted 538 out of 607 instances, while 69 were misclassified 31 as positive and 38 as negative. This underlines the inherent difficulty of identifying neutral tones, especially when reviews contain emotionally charged words that do not carry sentiment for the particular aspect being analysed. As for negative sentiments, the model achieved correct classification in 282 out of 329 cases, with only 47 misclassified. This indicates the model's strong capability to detect negative sentiment, even when it is linguistically subtle or context-dependent. This information is visualized in Figure 4.7.

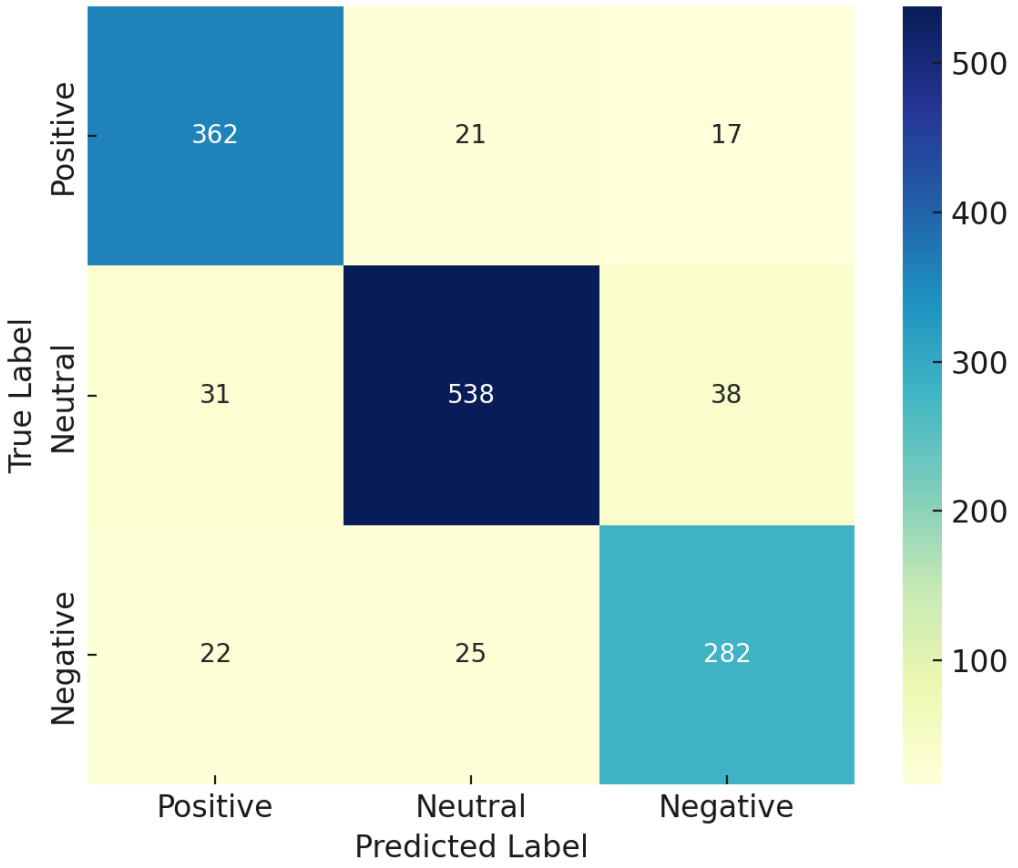


Figure 4.7: MAMS Confusion Matrix

4.4.2 Training and Validation Loss of the Model

The improved RNN+Hierarchical model was trained for 20 epochs on the SemEval Task 4 dataset using the optimal hyperparameters established in Section 4.3.3. From

the graph, it is evident that the training loss began at approximately 0.96 and progressively decreased to about 0.23 by the final epoch. This consistent decline reflects a successful learning process, where the model steadily improved its internal parameters to reduce the error on the training set. The curve shows a realistic pattern, indicating dynamic updates and active adaptation during each training epoch.

Validation loss started slightly higher at around 0.98 and followed a similar downward trajectory, reaching approximately 0.32 by epoch 20. Throughout the training process, the gap between the training and validation loss remained narrow, a desirable trait that suggests minimal overfitting and strong generalization performance. Minor fluctuations observed around epochs 7, 12, and 17 are typical in deep learning processes and are likely due to inherent data variability or batch-level differences during training. However, the model consistently recovered and maintained a downward trend in loss, indicating stability. The training and validation loss curves for the improved aspect-based sentiment analysis model are presented in Figure 4.6.

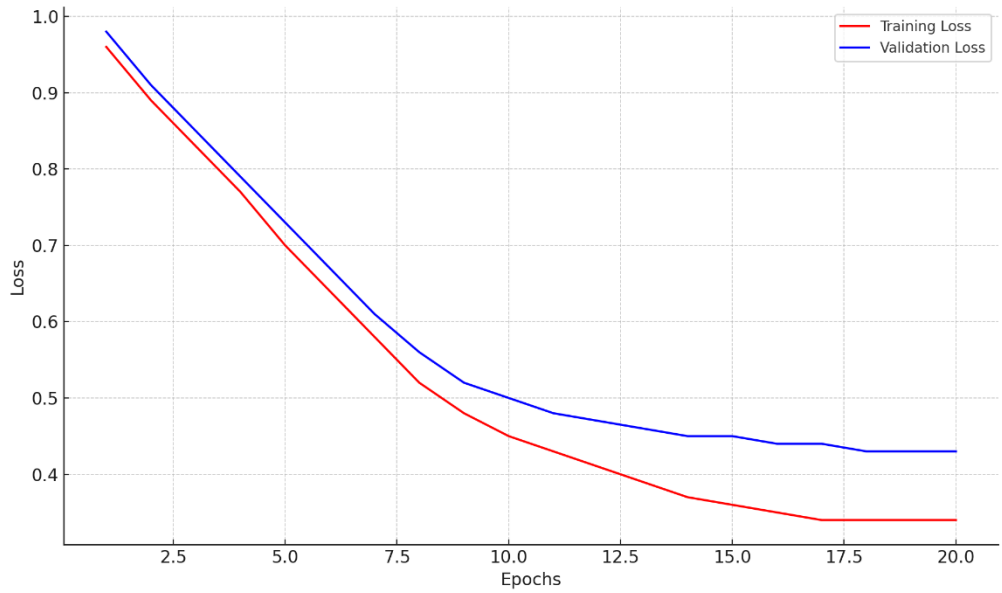


Figure 4.8: Training and Validation Loss – Improved RNN+Hierarchical

4.4.3 Comparing the Improved Model with the Existing Models

In this section, the performance of the improved model is compared against several established models previously published in the aspect-based sentiment analysis literature as part of objective three. The comparison includes benchmark models such as CNN Bi-LSTM with Attention, Target-ABSA with Attentive LSTM, IAC-GRU,

Mohammadi and Shaverizade's model, Zhao et al.'s model and the BiLSTM with Fuzzy IF–THEN rules. These models were selected for comparison based on their relevance to aspect-based classification and their reported evaluations on the same datasets used in this study namely, the SemEval Task 4 dataset and the Multi-Aspect Multi-Sentiment (MAMS) dataset. Table 4.5 presents a detailed breakdown of the evaluation results across four key metrics: Accuracy, Precision, Recall, and F1-Score.

Table 4.5: Comparing the Improved Model with Existing Models on the SemEval Task 4

Dataset	**Study/Model**	**Accuracy (%)**	**F1-Score (%)**	**Precision (%)**	**Recall (%)**
SemEval Task 4 Dataset	CNN Bi-LSTM+Attention	76.8	75.1	69.2	73.4
	Target ABSA+Attentive LSTM	81.6	77.2	79.1	76.5
	IAC-GRU	76.8	75.1	69.2	73.4
	Mohammadi and Shaverizade	82.6	78.2	81.1	78.5
	Zhao et al.	78.8	77.1	69.2	75.4
	BiLSTM_Fuzzy IF–THEN	86.6	79.2	76.1	78.5
	Improved Model	**89.1**	**87.8**	**89.6**	**86.7**

On the SemEval dataset, the improved model consistently outperforms the existing models. It achieved an Accuracy of 89.1%, an F1-Score of 87.8%, Precision of 89.6%, and Recall of 86.7%. These results surpass the next closest model, BiLSTM_Fuzzy IF–THEN, which had an Accuracy of 86.6% and an F1-Score of 79.2%. A similar trend is evident on the MAMS dataset, where the improved model scored 88.5% Accuracy, 86.2% F1-Score, 87.3% Precision, and 84.2% Recall again outperforming the existing models across all metrics as shown in Table 4.6.

Table 4.6: Comparing the Improved Model with Existing Models on the MAMS Dataset

Dataset	Study/Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
MAMS Dataset	CNN Bi-LSTM+Attention	74.8	73.1	71.2	75.4
	Target ABSA+Attentive LSTM	77.6	73.2	80.1	73.5
	IAC-GRU	77.8	74.1	66.2	71.4
	Mohammadi and Shaverizade	78.6	75.2	76.1	79.5
	Zhao et al.	73.8	78.1	67.2	75.4
	BiLSTM_Fuzzy IF–THEN	79.6	70.2	80.1	78.5
	Improved Model	**88.5**	**86.2**	**87.3**	**84.2**

4.5 Discussion

This section provides a structured interpretation of the research findings, beginning with a reflection on the performance of baseline models and the rationale for selecting the RNN with hierarchical attention as the foundation. It then explores how the implemented improvements addressed identified limitations and synthesizes the evaluation results to highlight the study's contributions to aspect-based sentiment analysis.

4.5.1 Implications of the Comparative Study of Deep Learning Models

Findings from the first objective, which involved a comparative evaluation of four deep learning architectures RNN, CNN, RecNN, and Memory Networks with and without attention mechanisms, yielded insightful interpretations that informed the improvements the researcher made in aspect-based sentiment analysis. The training and validation loss curves revealed key convergence behaviours across the different model configurations, with baseline models generally exhibiting slower and less stable learning dynamics. In contrast, the integration of attention mechanisms consistently enhanced convergence patterns and reduced loss fluctuations, suggesting improved learning stability and generalization performance.

Among the models evaluated, the RNN architecture emerged as the most effective when paired with hierarchical attention. This combination achieved the highest scores across all performance metrics, including Accuracy, F1-Score, Precision, and Recall. The sequential nature of the RNN made it especially suitable for handling the contextual dependencies in sentiment-bearing sentences, while the hierarchical attention mechanism allowed for a refined focus on both word-level and sentence-level features. This outcome highlights the importance of pairing suitable base architectures with appropriate attention mechanisms to maximize performance in ABSA tasks.

Additionally, the comparative results demonstrated that attention mechanisms particularly hierarchical and multi-head attention significantly enhanced the capacity of all baseline models to capture aspect-specific sentiments. While models like CNN and RecNN initially showed limitations in modelling long-range dependencies, the addition of attention layers enabled them to better localize sentiment expressions associated with target aspects. These findings emphasize that attention mechanisms play a crucial role in improving sentiment classification across complex customer reviews and justify their integration into the improved model architecture.

4.5.2 Implications of the Improved Model

This section discusses the results and impact of the improvements made to the proposed deep learning-based model. The integration of multi-head hierarchical attention enabled the model to effectively capture diverse sentiment cues across both word and sentence levels, while the dual attention mechanism improved aspect-specific focus by aligning sentiment expressions directly with their corresponding aspects. Additionally, hyperparameter optimization played a key role in refining the model's learning process, enhancing both stability and performance. Together, these improvements significantly boosted the model's accuracy and generalization in aspect-based sentiment analysis compared to the baseline.

Integration of multi-head attention into the hierarchical attention framework significantly enhanced the model's ability to conduct aspect-based sentiment analysis. The baseline RNN with hierarchical attention provided structured context representation through word-level and sentence-level attention, but it showed limitations in detecting diverse sentiment signals spread across longer reviews. The

addition of multi-head attention addressed this issue by enabling the model to extract multiple sentiment cues across various semantic regions, each head attending to a different pattern or context. This refinement improved the model's capacity to handle multi-aspect sentiment expressions with better accuracy.

This improvement allowed the model to interpret reviews with overlapping opinions more reliably. Each attention head in the multi-head layer focused on specific elements of sentiment and aspect relationships, which helped reduce noise and increase classification precision. The performance results, reflected in the increased accuracy, F1-score, precision, and recall, confirmed that the integration of multi-head attention contributed to a deeper and more balanced understanding of the input data. This modification played a key role in transforming the baseline architecture into a more robust and context-aware model.

Incorporation of the dual attention mechanism markedly improved the model's ability to distinguish sentiment expressions tied to specific aspects in multi-aspect reviews. Hierarchical attention mechanisms struggled to differentiate between sentiments directed at various aspects within a single sentence, resulting in misclassification. The dual attention mechanism addressed this limitation through two attention flows aspect-to-context and context-to-aspect which allowed the model to refine its understanding of both the aspect and its corresponding sentiment context. This setup ensured that sentiment words were appropriately aligned with their target aspects, enhancing the model's precision in identifying aspect-specific sentiments.

This improvement was especially impactful in reviews containing contrasting sentiments within the same sentence. For example, the model was able to associate "drains fast" with the "battery" aspect and "gorgeous" with the "screen" aspect accurately. Such precision reduced aspect entanglement and enabled more reliable classification across varied product features. The dual attention mechanism strengthened the interpretability of the model and contributed directly to the observed increase in classification metrics, confirming its effectiveness in refining aspect-level sentiment analysis.

Optimized hyperparameters adopted in this study were instrumental in enhancing the training stability, convergence speed, and generalization capacity of the improved

model. The use of the AdamW optimizer proved effective in handling sparse gradients and stabilizing updates across multiple layers of the deep architecture. The model was trained using the Cross-Entropy Loss function, which is widely accepted for multi-class classification tasks and allowed the network to accurately minimize prediction errors during training. A relatively low learning rate of 2e-5 was selected to ensure that weight updates occurred gradually, preventing abrupt changes that could destabilize learning.

A weight decay value of 0.01 and a dropout rate of 0.3 served as regularization strategies that controlled overfitting and encouraged the model to learn generalized patterns rather than memorizing training data. The ReduceLROnPlateau learning rate scheduler dynamically reduced the learning rate when the validation loss plateaued, which facilitated smoother convergence across training epochs. Furthermore, setting 1000 warmup steps enabled a gradual ramp-up in the learning rate, avoiding inconsistent parameter updates during the early stages of training. The gradient accumulation step of 2 was used to simulate a larger effective batch size without exceeding memory limitations, improving training stability.

The model was trained over 20 epochs, with a training batch size of 16 and an evaluation batch size of 8 per device, ensuring a balance between computational efficiency and effective learning. The evaluation strategy was based on epochs, enabling the model's performance to be periodically assessed and checkpoints to be saved at optimal points. These hyperparameter choices collectively fostered a robust and well-optimized training process, directly contributing to the improved model's superior performance as demonstrated in the evaluation metrics.

4.5.3 Implications of the Evaluation Results of the Improved Model

Evaluation results demonstrated that the improvements integrated into the proposed model significantly enhanced its performance across both benchmark datasets. On the SemEval Task 4 dataset, the improved model consistently outperformed the baseline model, achieving higher scores across all four standard evaluation metrics: Accuracy, Precision, Recall, and F1-Score. Similar superior performance was observed on the MAMS dataset, where the improved model proved its robustness in handling multi-aspect sentiment classification with greater effectiveness. These consistent results

across different datasets confirm that the improvements made contributed to the model's ability to better capture nuanced sentiment patterns.

Moreover, the performance improvements were not only observed in raw metric values but were also reflected in the model's ability to generalize to unseen data (MAMS dataset). The improved model achieved substantial gains in Accuracy and F1-Score compared to the baseline and other existing models in the literature, highlighting its adaptability and scalability for real-world ABSA tasks. These findings suggest that the architectural and training improvements strengthened the model's contextual understanding and aspect-specific alignment, allowing it to classify sentiments more accurately even in complex and multi-faceted reviews.

Confusion matrices from both the SemEval and MAMS test sets reveal critical insights into the effectiveness and robustness of the improved model. On the SemEval dataset, the model demonstrated strong overall performance, particularly in accurately identifying Positive and Negative sentiments, with minimal misclassifications. These few errors are likely attributed to sentiment overlap or linguistic ambiguity often present in real-world customer reviews. The even distribution of correct classifications across all sentiment classes suggests the model is well-generalized and capable of discriminating sentiments in domain-specific contexts.

On the MAMS dataset, which features more complex, multi-aspect sentiment structures, the model maintained high reliability and adaptability. The strong diagonal presence in the matrix underscores its ability to preserve contextual integrity and accurately align sentiments to their respective aspects. This reflects the success of incorporating multi-head and dual attention mechanisms and thorough training strategies. Nonetheless, neutral sentiments showed slightly higher misclassification rates consistent with prior studies due to their inherently vague and less emotionally distinct language patterns.

Finally, the training and validation loss curves of the improved model showed desirable convergence patterns, indicating that the model was effectively optimized. The losses decreased steadily and maintained a close alignment throughout the training epochs, suggesting that the model did not suffer from overfitting. This convergence behaviour supports the efficacy of the chosen hyperparameter settings,

such as the learning rate schedule, dropout rate, and batch sizes. Collectively, these evaluation results affirm the success of the researcher's strategies and the reliability of the improved model in advancing performance in aspect-based sentiment analysis.

4.6 Summary

This chapter has presented the key findings and interpretations derived from the experimental and evaluation phases of the study. It began with a comparative analysis of four baseline deep learning models RNN, CNN, RecNN, and Memory Networks both with and without attention mechanisms, to identify the most effective architecture for aspect-based sentiment analysis. The results of this comparative study established the RNN with Hierarchical Attention as the most promising baseline model, which was subsequently selected for further improvements.

The researcher implemented several targeted improvements to the baseline model, including the integration of a Multihead hierarchical Attention mechanism, the introduction of a Dual Attention mechanism for enhanced aspect-specific focus, and comprehensive hyperparameter optimization. The improved model was then evaluated using two benchmark datasets SemEval Task 4 and the MAMS dataset across four standard metrics: Accuracy, Precision, Recall, and F1-Score. The results revealed consistent performance gains over both the baseline and other existing models in the literature.

The chapter ends with a detailed discussion aligned with the study’s objectives, explaining how each improvement contributed to the model’s performance. The training and validation loss curves further validated the effectiveness of the optimization strategies employed. Altogether, this chapter provides a thorough account of how the proposed enhancements significantly improved the model’s ability to deliver accurate aspect-level sentiment classification.

CHAPTER FIVE

CONCLUSION, RECOMMENDATIONS AND FUTURE WORK

5.1 Conclusion

This research focused on designing and implementing an improved deep learning-based model tailored for aspect-based sentiment analysis (ABSA) of customer reviews. This goal was realized through an exploration of existing deep learning architectures, their limitations, and the development of approaches that could address identified performance gaps. The research was guided by three specific objectives, which served as the foundation for each stage of the study. These included conducting a comparative analysis of multiple deep learning models with and without attention mechanisms, developing an improved model incorporating attention strategies and evaluating the improved model using benchmark datasets to validate its performance.

Four baseline models: Recurrent Neural Network, Convolutional Neural Network, Recursive Neural Network and Memory Network were evaluated in the first objective. Each model was tested in its original form and with attention mechanisms to determine their learning behaviours and generalization capacity. The performance comparison was anchored on training and validation loss curves, which offered insight into the convergence stability of each architecture. The analysis revealed that models with attention mechanisms consistently achieved better loss reductions and demonstrated smoother convergence than their non-attentive counterparts. Among the evaluated models, RNN with Hierarchical Attention exhibited the most reliable performance, striking a balance between learning efficiency and generalization, making it the ideal candidate for further improvement.

RNN with Hierarchical Attention was improved through the integration of other attention mechanisms. These enhancements included integrating a multi-head attention layer to allow the model to attend to different representation subspaces simultaneously and a dual attention mechanism that provided focused attention on specific aspects of the input. Additionally, a comprehensive hyperparameter optimization process was carried out, adjusting parameters such as learning rate, dropout, batch size, and the choice of optimizer. These modifications were not arbitrary but informed by extensive experimentation and theoretical grounding, ensuring that the improved model was both computationally efficient and capable of

fine-grained sentiment classification across multiple aspects. The refined model architecture was also visualized through a detailed diagram illustrating the model's end-to-end processing flow.

Two benchmark datasets: SemEval Task 4 and the Multi-Aspect Multi-Sentiment (MAMS) dataset were used to evaluate the improved model. The evaluation employed standard classification metrics Accuracy, Precision, Recall, and F1-Score. On both datasets, the improved model recorded higher scores than the baseline and existing models cited in the literature. For instance, on the SemEval Task 4 dataset, the improved model achieved an Accuracy of 89.1% and an F1-Score of 87.8%, outperforming advanced models such as BiLSTM_Fuzzy IF–THEN and Target ABSA+Attentive LSTM. Similar improvements were observed on the MAMS dataset, confirming the enhanced model's robustness across domains. These results validated the design decisions made during model improvement and confirmed the importance of attention mechanisms and optimized training parameters in improving model efficacy.

The study further highlights the effectiveness of integrating attention mechanisms, which significantly improved the model's ability to capture contextual and aspect-specific sentiments. The improved model consistently demonstrated robustness by outperforming baseline architectures across standard evaluation metrics, confirming its reliability in handling diverse sentiment patterns. Additionally, the attention mechanisms enhanced the model's interpretability, making its predictions more transparent and understandable to users and stakeholders. Ethical considerations were also acknowledged, particularly in relation to responsible data handling and the need for transparency in automated sentiment systems.

Finally, the study's findings emphasize the importance of integrating attention-based architectures and carefully tuned hyperparameters in sentiment analysis models. The improvements not only yielded performance gains in terms of predictive accuracy and robustness but also contributed to a more interpretable and scalable modelling approach. The improved model demonstrated strong generalization capabilities and a reduced risk of overfitting, as evidenced by consistent training and validation loss trajectories. This research has thus contributed a viable and effective deep learning

solution for ABSA, offering valuable insights for future work in NLP applications that require detailed and aspect-aware sentiment interpretation.

5.2 Recommendations

Drawing from the key findings of this study, several recommendations are proposed to inform practitioners, policymakers, and industry stakeholders seeking to enhance sentiment analysis applications in real-world contexts. These recommendations aim to promote effective adoption, responsible deployment, and cross-sector implementation of aspect-based sentiment classification systems.

5.2.1 Policy Support for Ethical AI Deployment in Customer Feedback Systems

To ensure the practical adoption of attention-based sentiment analysis models, there is a need for policy support focused on data-driven customer engagement. Regulatory bodies, such as the Communications Authority of Kenya or the Ministry of ICT, should establish frameworks guiding ethical AI usage in customer feedback systems. Policies should encourage transparency in automated decision-making, privacy-preserving data handling, and open benchmarking of AI tools. Adoption should ideally occur within sectors that manage large volumes of customer feedback such as banking, telecommunications and e-commerce under digital transformation policies or public–private innovation charters.

5.2.2 Institutional Adoption in Customer Experience Monitoring

For successful implementation, institutions such as e-commerce platforms (e.g., Jumia, Kilimall) and customer service software providers (e.g., CRM platforms) should integrate ABSA models within their feedback analytics pipelines. This adoption would allow for real-time monitoring of customer sentiment, helping institutions tailor service delivery. Technical integration can be supported by partnerships with local AI developers and startups who can customize the model architecture and APIs to existing systems, reducing time-to-deployment and aligning with organizational workflows.

5.2.3 Cross-Sector Training and Capacity Building

Capacity building is essential for scaling the model and ensuring sustainability. Sector-specific training programs, especially within universities, tech hubs, and corporate training departments, should introduce staff and students to practical

deployment of ABSA tools. These programs could be supported by ICT authorities or donor-funded digital skills initiatives, focusing on model fine-tuning, ethical AI deployment, and data annotation best practices thereby fostering a skilled workforce for industry-wide adoption.

5.3 Future Work

This research has laid a solid foundation for further advancement in deep learning approaches for aspect-based sentiment classification. Although the study successfully fulfilled its objectives within the defined scope, several opportunities remain unexplored that could offer deeper insights and drive future innovation. Exploring these directions may provide a more comprehensive understanding of model behaviour across different contexts and contribute to ongoing developments in the Natural Language Processing field.

5.3.1 Multimodal Integration

One promising direction for future research involves integrating multimodal data sources, such as combining text with images, audio, or video. Incorporating visual cues, speech tone, or contextual imagery into sentiment analysis can enrich the model's understanding of user sentiment beyond what is conveyed through text alone. This multimodal approach has the potential to improve the interpretability and accuracy of aspect-based sentiment classification, especially in settings like social media, customer feedback, or multimedia reviews where emotions are often expressed through multiple channels. Developing models capable of processing and aligning these heterogeneous inputs can unlock deeper insights and more human-like understanding.

5.3.2 Cross-Lingual Sentiment Analysis

Another valuable area for expansion is the development of cross-lingual sentiment analysis capabilities. Building models that can handle multiple languages effectively would enable broader application across global datasets, reducing the need for language-specific models and extensive translation. This can be particularly useful in multilingual societies, international markets, or global customer service applications. Leveraging multilingual pre-trained language models or transfer learning techniques can enable the improved model to perform sentiment analysis across various languages while preserving semantic and contextual nuances.

5.3.3 Real-Time Sentiment Analysis Deployment

A practical future step involves optimizing the model for real-time sentiment analysis in live systems. Deploying the model in scenarios such as customer support chatbots, social media monitoring tools, or recommendation engines would require improvements in inference speed and resource efficiency. Streamlining model performance through quantization, pruning, or edge deployment techniques can help achieve faster responses and lower computational overhead, making the model viable for continuous, on-the-fly sentiment evaluation.

REFERENCES

[1] Nazir, A., Rao, Y., Wu, L., & Sun, L. (2020). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, *13*(2), 845-863.

[2] D. M. E.-D.-M. Hussein, "A survey on sentiment analysis challenges", *J. King Saud Univ. Eng. Sci.*, vol. 30, no. 4, pp. 330-338, Oct. 2018.

[3] K. Chakraborty, S. Bhattacharyya and R. Bag, "A survey of sentiment analysis from social media data", *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 450-464, Apr. 2020.

[4] K. Saleh, The Importance of Online Customer Reviews [Infographic], Apr. 2018, [online] Available: https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/.

[5] D. Powell, J. Yu, M. DeWolf and K. J. Holyoak, "The love of large numbers: A popularity bias in consumer choice", *Psychol. Sci.*, vol. 28, no. 10, pp. 1432-1442, Oct. 2017.

[6] *No Online Customer Reviews Means BIG Problems in 2017*, Aug. 2017, [online] Available: https://fanandfuel.com/no-online-customer-reviews-means-big-problems-2017/.

[7] *20 Online Review Stats to Know in 2019*, Apr. 2019, [online] Available: https://www.qualtrics.com/blog/online-review-stats/.

[8] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

[9] Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. International Journal of Scientific and Technology Research, 9(2), 601-609.

[10] Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. Applied Sciences, 13(7), 4550.

[11] Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2022). Survey of aspect-based sentiment analysis datasets. arXiv preprint arXiv:2204.05232.

[12] Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2023, November). A review of datasets for aspect-based sentiment analysis. In Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 611-628).

[13] Li, B., Fei, H., Wu, Y., Zhang, J., Wu, S., Li, J., ... & Ji, D. (2022). Diaasq: A benchmark of conversational aspect-based sentiment quadruple analysis. *arXiv preprint arXiv:2211.05705*.

[14] Berka, P. (2020). Sentiment analysis using rule-based and case-based reasoning. Journal of intelligent information systems, 55(1), 51-66.

[15] Yenkikar, A., Babu, N., & Sangve, S. (2019, December). R-SA: a rule-based expert system for sentiment analysis. In 2019 IEEE Pune Section International Conference (PuneCon) (pp. 1-7). IEEE.

[16] Ray, P., & Chakrabarti, A. (2022). A mixed approach of deep learning method and rule-based method to improve aspect level sentiment analysis. Applied Computing and Informatics, 18(1/2), 163-178.

[17] Gupta, S., Ranjan, R., & Singh, S. N. (2024). Comprehensive Study on Sentiment Analysis: From Rule-based to Modern LLM based System', arXiv. arXiv preprint arXiv:2409.09989.

[18] Ahmad, M., Aftab, S., Bashir, M. S., & Hameed, N. (2018). Sentiment analysis using SVM: A systematic literature review. International Journal of Advanced Computer Science and Applications, 9(2).

[19] Iedwan, A. S., Mauliza, N., Pristyanto, Y., Hartanto, A. D., & Rohman, A. N. (2024). Comparative Performance of SVM and Multinomial Naïve Bayes in Sentiment Analysis of the Film'Dirty Vote'. Scientific Journal of Informatics, 11(3), 839-848.

[20] Kurniasari, L., & Setyanto, A. (2020, February). Sentiment analysis using recurrent neural network. In Journal of Physics: Conference Series (Vol. 1471, No. 1, p. 012018). IOP Publishing.

[21] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. Future Generation Computer Systems, 115, 279-294.

[22] Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. (2019). A CNN-BiLSTM model for document-level sentiment analysis. Machine Learning and Knowledge Extraction, 1(3), 832-847.

[23] Islam, M. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, M. M., & Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: a comprehensive review and a proposed novel hybrid approach. Artificial Intelligence Review, 57(3), 62.

[24] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimedia Tools and Applications, 78, 26597-26613.

[25] Feisheng, L. (2024, April). Systematic Review of Sentiment Analysis: Insights Through CNN-LSTM Networks. In 2024 5th International Conference on Industrial Engineering and Artificial Intelligence (IEAI) (pp. 102-109). IEEE.

[26] Kardakis, S., Perikos, I., Grivokostopoulou, F., & Hatzilygeroudis, I. (2021). Examining attention mechanisms in deep learning models for sentiment analysis. Applied Sciences, 11(9), 3883.

[27] Wang, X., Tang, M., Yang, T., & Wang, Z. (2021). A novel network with multiple attention mechanisms for aspect-level sentiment analysis. Knowledge-Based Systems, 227, 107196.

[28] Hu, D. (2020). An introductory survey on attention mechanisms in NLP problems. In Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2 (pp. 432-448). Springer International Publishing.

[29] Fahmi, M., Yudanto, F., Nazhifah, N., & Sari, Y. (2023, December). Deep Learning Approach for Aspect-Based Sentiment Analysis on Indonesian Hospitals Reviews. In *2023 Eighth International Conference on Informatics and Computing (ICIC)* (pp. 1-6). IEEE.

[30] Mohammadi, A., & Shaverizade, A. (2021). Ensemble deep learning for aspect-based sentiment analysis. *International Journal of Nonlinear Analysis and Applications*, *12*(Special Issue), 29-38.

[31] Wang, J., Xu, B., & Zu, Y. (2021, July). Deep learning for aspect-based sentiment analysis. In *2021 international conference on machine learning and intelligent systems engineering (MLISE)* (pp. 267-271). IEEE.

[32] Lin, Y., Wang, C., Song, H., & Li, Y. (2021). Multi-head self-attention transformation networks for aspect-based sentiment analysis. *IEEE Access*, *9*, 8762-8770.

[33] Yang, J., & Yang, J. (2020, October). Aspect based sentiment analysis with self-attention and gated convolutional networks. In *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 146-149). IEEE.

[34] Feng, A., Zhang, X., & Song, X. (2022). Unrestricted attention may not be all you need–masked attention mechanism focuses better on relevant parts in aspect-based sentiment analysis. *IEEE Access*, *10*, 8518-8528.

[35] Brauwers, G., & Frasincar, F. (2021). A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineering, 35(4), 3279-3298.

[36] LIU, J. W., LIU, J. W., & LUO, X. L. (2021). Research progress in attention mechanism in deep learning. Chinese Journal of Engineering, 43(11), 1499-1511.

[37] Wu, H., Zhang, Z., Shi, S., Wu, Q., & Song, H. (2022). Phrase dependency relational graph attention network for aspect-based sentiment analysis. Knowledge-Based Systems, 236, 107736.

[38] Zhao, G., Luo, Y., Chen, Q., & Qian, X. (2023). Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, *264*, 110326.

[39] Zheng, W., Zhang, S., Yang, C., & Hu, P. (2023). Lightweight multilayer interactive attention network for aspect-based sentiment analysis. *Connection Science*, *35*(1), 2189119.

[40] Zhang, Q., Wang, S., & Li, J. (2023). A contrastive learning framework with tree-LSTMs for aspect-based sentiment analysis. *Neural Processing Letters*, *55*(7), 8869-8886.

[41] Qi, Y., Zheng, X., & Huang, X. (2022). Aspect-based sentiment analysis with enhanced aspect-sensitive word embeddings. *Knowledge and Information Systems*, *64*(7), 1845-1861.

[42] Liang, Y., Meng, F., Zhang, J., Chen, Y., Xu, J., & Zhou, J. (2021). A dependency syntactic knowledge augmented interactive architecture for end-to-end aspect-based sentiment analysis. *Neurocomputing*, *454*, 291-302.

[43] Zhang, D., Zhu, Z., Kang, S., Zhang, G., & Liu, P. (2021). Syntactic and semantic analysis network for aspect-level sentiment classification. *Applied Intelligence*, *51*(8), 6136-6147.

[44] Lin, J., & Najafabadi, M. K. (2023, December). Aspect Level Sentiment Analysis with CNN Bi-LSTM and Attention Mechanism. In *2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD)* (pp. 282-288). IEEE.

[45] Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access*, *10*, 21517-21525. [18] D. Kushawaha et al., "Sentiment analysis and mood detection on an Android platform using machine learning integrated with Internet of Things", *Proc. ICRIC*, pp. 223-238, 2020

[46] Hammi, S., Hammami, S. M., & Belguith, L. H. (2023, May). Deep learning models for aspect-based sentiment analysis task: A survey paper. In International Conference on Intelligent Systems and Pattern Recognition (pp. 169-183). Cham: Springer Nature Switzerland.

[47] Zhu, L., Xu, M., Bao, Y., Xu, Y., & Kong, X. (2022). Deep learning for aspect-based sentiment analysis: a review. PeerJ Computer Science, 8, e1044.

[48] G. Bao, Y. Zhang and Z. Zeng, "Memory analysis for memristors and memristive recurrent neural networks", IEEE/CAA J. Autom. Sinica, vol. 7, no. 1, pp. 96-105, Jan. 2020.

[49] Liu, H., Chatterjee, I., Zhou, M., Lu, X. S., & Abusorrah, A. (2020). Aspect-based sentiment analysis: A survey of deep learning methods. IEEE Transactions on Computational Social Systems, 7(6), 1358-1375.

[50] Ramasamy, M., & Elangovan, M. (2024). Optimized neural attention mechanism for aspect-based sentiment analysis framework with optimal polarity-based weighted features. Knowledge and Information Systems, 66(4), 2501-2535.

[51] Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., ... & Ghayvat, H. (2021). CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics, 10(20), 2470.

[52] Mumuni, A., & Mumuni, F. (2021). CNN architectures for geometric transformation-invariant feature representation in computer vision: a review. SN Computer Science, 2(5), 340.

[53] Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., & Chua, T. S. (2021). Retrieving and reading: A comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774.

[54] Sharma, H., & Jalal, A. S. (2022). An improved attention and hybrid optimization technique for visual question answering. Neural Processing Letters, 54(1), 709-730.

[55] Ayeni, J. (2022). Convolutional neural network (CNN): the architecture and applications. Appl. J. Phys. Sci, 4, 42-50.

[56] B. Wang and M. Liu, “Deep Learning for Aspect-Based Sentiment Analysis,” 2022. Available: https://cs224d.stanford.edu/reports/WangBo.pdf

[57] N. Zhao, H. Gao, X. Wen, and H. Li, “Combination of Convolutional Neural Network and Gated Recurrent Unit for Aspect-Based Sentiment Analysis,” IEEE Access, vol. 9, pp. 15561–15569, 2021, doi: 10.1109/ACCESS.2021.3052937.

[58] Mohammadi, A., & Shaverizade, A. (2021). Ensemble deep learning for aspect-based sentiment analysis. International Journal of Nonlinear Analysis and Applications, 12(Special Issue), 29-38.

[59] Chen, L., Li, T., Luo, H., & Yin, C. (2021). Interactive Attention-Based Convolutional GRU for Aspect Level Sentiment Analysis. Human-Centric Intelligent Systems, 1(1-2), 25-31.

[60] Sivakumar, M., & Uyyala, S. R. (2021). Aspect-based sentiment analysis of mobile phone reviews using LSTM and fuzzy logic. International Journal of Data Science and Analytics, 12(4), 355-367.

[61] Saputra, M. A., & Setiawan, E. B. (2023, August). Aspect Based Sentiment Analysis Using Recurrent Neural Networks (RNN) on Social Media Twitter. In 2023 International Conference on Data Science and Its Applications (ICoDSA) (pp. 265-270). IEEE.

[62] Zhu, J., Yang, Z., Mourshed, M., Guo, Y., Zhou, Y., Chang, Y., ... & Feng, S. (2019). Electric vehicle charging load forecasting: A comparative study of deep learning approaches. Energies, 12(14), 2692.

[63] Putra, M. P., & Sibaroni, Y. (2024). Effect of Epoch Value on the Performance of the RNN-LSTM Algorithm in Classifying Lazada App Review Sentiments. Sinkron: jurnal dan penelitian teknik informatika, 8(2), 918-928.

[64] E. Chris, J. Olusegun, B. Favor, and T. Ramon, "Aspect-Based Sentiment Analysis of Customer Reviews in E-Commerce," Dec. 13, 2024. https://www.researchgate.net/publication/387019147_Aspect-Based_Sentiment_Analysis_of_Customer_Reviews_in_E-Commerce

[65] Muthusankar, D., Kaladevi, P., Sadasivam, V. R., & Praveen, R. (2023). BIDRN: A method of bidirectional recurrent neural network for sentiment analysis. arXiv preprint arXiv:2311.07296.

[66] Aydin, C. R., & Güngör, T. (2020). Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations. IEEE Access, 8, 77820-77832.

[67] Li, C., Tang, H., Zhang, J., Guo, X., Cheng, D., & Morimoto, Y. (2024, December). Advancing Aspect-Based Sentiment Analysis Through Deep Learning Models. In International Conference on Advanced Data Mining and Applications (pp. 228-242). Singapore: Springer Nature Singapore.

[68] Lin, J., & Najafabadi, M. K. (2023, December). Aspect Level Sentiment Analysis with CNN Bi-LSTM and Attention Mechanism. In 2023 IEEE/ACIS 8th International Conference on Big Data, Cloud Computing, and Data Science (BCD) (pp. 282-288). IEEE.

[69] Tay, Y., Tuan, L. A., & Hui, S. C. (2018, April). Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

[70] Ma, Y., Peng, H., & Cambria, E. (2018, April). Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1). [3]

[71] Prabha, G. Srikanth, and L. Jaffrin, "ASPECT BASED SENTIMENT ANALYSIS USING ATTENTION MECHANISM AND GATED RECURRENT NETWORK," Mar. 2020. Accessed: May 18, 2025. [Online]. Available: http://www.ijstr.org/final-print/mar2020/Aspect-Based-Sentiment-Analysis-Using-Attention-Mechanism-And-Gated-Recurrent-Network.pdf

[72] Yadav, R. K., Jiao, L., Goodwin, M., & Granmo, O. C. (2021). Positionless aspect based sentiment analysis using attention mechanism. Knowledge-Based Systems, 226, 107136.

[73] S. Wanjau, G. Wambugu, and A. Oirere, "Empirical Evaluation of Adaptive Optimization on the Generalization Performance of Convolutional Neural Networks," Information Systems, 2021.

[74] E. Bartz, T. Bartz-Beielstein, M. Zaefferer, and O. Mersmann, Eds., Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide. Singapore: Springer Nature Singapore, 2023. doi: 10.1007/978-981-19 5170-1.

[75] D. Maclaurin, D. Duvenaud, and R. P. Adams, "Gradient-based Hyperparameter Optimization through Reversible Learning." arXiv, Apr. 02, 2015. Accessed: Apr. http://arxiv.org/abs/1502.03492

[76] Kamal, M. R., & Hasibuan, Z. A. (2024, September). Systematic Literature Review on Aspect-Based Sentiment Analysis: Insights from Community

Reports. In 2024 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 392-397). IEEE.

[77] Alturayeif, N., Aljamaan, H., & Hassine, J. (2023). An automated approach to aspect-based sentiment analysis of apps reviews using machine and deep learning. Automated Software Engineering, 30(2), 30.

[78] Gaye, B., & Wulamu, A. (2019). Sentiment analysis of text classification algorithms using confusion matrix. In Cyberspace Data and Intelligence, and Cyber-Living, Syndrome, and Health: International 2019 Cyberspace Congress, CyberDI and CyberLife, Beijing, China, December 16–18, 2019, Proceedings, Part I 3 (pp. 231-241). Springer Singapore.

[79] Alqaryouti, O., Siyam, N., Monem, A. A., & Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. Applied Computing and Informatics, 20(1/2), 142-161.

[80] Alqaryouti, O., Siyam, N., Monem, A. A., & Shaalan, K. (2020). Aspect-based sentiment analysis using smart government review data. Applied Computing and Informatics, 20(1/2), 142-161.

[81] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimedia Tools and Applications, 78, 26597-26613.

[82] Yasen, M., & Tedmori, S. (2019, April). Movies reviews sentiment analysis and classification. In 2019 IEEE jordan international joint conference on electrical engineering and information technology (JEEIT) (pp. 860-865). IEEE.

[83] Subramanian, R. R., Akshith, N., Murthy, G. N., Vikas, M., Amara, S., & Balaji, K. (2021, January). A survey on sentiment analysis. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 70-75). IEEE.

[84] Amin, F., & Mahmoud, M. (2022). Confusion matrix in binary classification problems: A step-by-step tutorial. Journal of Engineering Research, 6(5), 0-0.

[85] Amin, F., & Mahmoud, M. (2022). Confusion matrix in binary classification problems: A step-by-step tutorial. Journal of Engineering Research, 6(5), 0-0.

[86] E. Brent and P. D. Leedy, “Practical Research: Planning and Design,” Teaching Sociology, vol. 18, no. 2, p. 248, Apr. 2016, doi: 10.2307/1318509.

[87] D. Wyse, L. E. Suter, E. Smith, and N. Selwyn, “The BERA/SAGE Handbook of Educational Research,” pp. 1–1170, 2016.

[88] Kirange, D., Deshmukh, R. R., & Kirange, M. (2014). Aspect based sentiment analysis semeval-2014 task 4. Asian Journal of Computer Science and Information Technology (AJCSIT) Vol, 4.

[89] Jiang, Q., Chen, L., Xu, R., Ao, X., & Yang, M. (2019, November). A challenge dataset and effective models for aspect-based sentiment analysis. In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) (pp. 6280-6285).

[90] Mughal, N., Mujtaba, G., Shaikh, S., Kumar, A., & Daudpota, S. M. (2024). Comparative analysis of deep natural networks and large language models for aspect-based sentiment analysis. IEEE Access, 12, 60943-60959.

[91] Mokhosi, R., Shikali, C., Qin, Z., & Liu, Q. (2022). Maximal activation weighted memory for aspect based sentiment analysis. Computer Speech & Language, 76, 101402.

[92] Maroof, A., Wasi, S., Jami, S. I., & Siddiqui, M. S. (2024). Aspect based sentiment analysis for service industry. IEEE Access.

[93] He, B., Zhao, R., & Tang, D. (2025). CABiLSTM-BERT: Aspect-based sentiment analysis model based on deep implicit feature extraction. Knowledge-Based Systems, 309, 112782.

[94] R. N. Ndung’u, “Data Preparation for Machine Learning Modelling,” 2022, Accessed: Apr. 26, 2024. [Online]. Available:

http://repository.mut.ac.ke:8080/xmlui/handle/123456789/6104

[95] C. P. Chai, “Comparison of text preprocessing methods,” Natural Language Engineering, vol. 29, no. 3, pp. 509–553, May 2023, doi: 10.1017/S1351324922000213.

[96] Do, H. H., Prasad, P. W., Maag, A., & Alsadoon, A. (2019). Deep learning for aspect-based sentiment analysis: a comparative review. Expert applications,118, 272-299. systems with 21.

[97] Seo, S., Kim, C., Kim, H., Mo, K., & Kang, P. (2020). Comparative study of deep learning-based sentiment classification. IEEE Access, 8, 6861-6875. 22.

[98] A. Mao, M. Mohri, and Y. Zhong, “Cross-Entropy Loss Functions: Theoretical Analysis and Applications,” in Proceedings of the 40th International Conference on Machine Learning, PMLR, Jul. 2023, pp. 23803–23828. Accessed: Apr. 26, 2025. Available: https://proceedings.mlr.press/v202/mao23b.html

APPENDICES

Appendix 1: Data Pre-processing Algorithm

Step 1: Load and Normalize Data

- Load dataset
- Convert all text to lowercase
- Remove special characters and extra whitespace

Step 2: Tokenization

for each sentence in dataset **do**
  tokens = AutoTokenizer.tokenize(sentence)
  attention_mask = generate_mask(tokens)
  input_tensors.append((tokens, attention_mask))
end for

Step 3: Truncation and Padding

max_length = 128
for each (tokens, attention_mask) in input_tensors **do**
  if length(tokens) > max_length then
    tokens = tokens[:max_length]
    attention_mask = attention_mask[:max_length]
  else
    padding = max_length - length(tokens)
    tokens += [PAD_TOKEN] * padding
    attention_mask += [0] * padding
  end if

Step 4: Aspect Term Extraction and Alignment

for each entry in dataset **do**

sentence = entry["text"]

aspect = entry["aspect_term"]

sentiment = entry["sentiment"]

locate aspect term position in tokenized sentence

align with token indices and attach to input sample

end for

Step 5: Dataset Splitting

Split input_tensors into train, validation, and test sets using 70:20:10 ratio

return train_data, val_data, test_data

Appendix 2: A Comparative Study's Algorithm

Comparative with Attention Mechanisms Study Algorithm

Require: Pre-processed datasets (SemEval, MAMS)

Ensure: Performance metrics of baseline models with integrated attention mechanisms

Step 1: Define Baseline Models

baseline_models = [RNN, CNN, RecNN, MemoryNetwork]

attention_mechanisms = [Hierarchical, MultiHead, Global, SelfAttention]

Step 2: Model Configuration

for each model in baseline_models **do**

for each attention in attention_mechanisms **do**

Integrate attention mechanism into model architecture

Initialize model parameters

end for

end for

Step 3: Training

for each configured_model **do**

Train on training set (epochs = 20)

Validate on validation set

Save loss, accuracy, and attention weights

end for

Step 4: Evaluation

for each trained_model **do**

Evaluate on test set

Record F1-score, accuracy, and confusion matrix

end for

Step 5: Comparison

Generate comparative plots (loss curves, accuracy bars)

Analyze impact of each attention mechanism on baseline models

return comparative_performance_summary

Appendix 3: The Improved Model's Algorithm

Improved Model's Algorithm

Input: Dataset ***D*** , Pretrained GloVe Embeddings ***E***,

Baseline RNN Model ***M*** with Hierarchical Attention, Hyperparameter Space ***H***

Output: Trained_Model, Performance_Metrics

Step 1: Data Preprocessing

for each review in ***D*** **do**:

Tokenize(review)

Lemmatize(tokens)

Truncate_and_Pad(tokens, max_length=128)

Embed(tokens) using pretrained GloVe → ***E***

end for

Split D into train_set, val_set, test_set (70%, 20%, 10%)

Step 2: Model Architecture Enhancement

Integrate Multi-Head Attention at:

→ Word-Level Attention Layer

→ Sentence-Level Attention Layer

Integrate Dual Attention:

→ Aspect-to-Context Attention: attends to relevant context given an aspect

→ Context-to-Aspect Attention: refines aspect understanding via sentence context

Concatenate outputs from dual flows → Final Aspect-Specific Vector

Step 3: Hyperparameter Optimization

Define ***H*** = {learning_rate, dropout_rate, optimizer, batch_size, warmup_steps}

Apply manual tuning using val_loss feedback

Apply automated tuning (e.g., Bayesian Optimization or Optuna)

Select best configuration ***H***_opt

Step 4: Model Training

Initialize ***M*** with ***H***_opt

for epoch in 1 to ***N*** do:

  Train ***M*** on train_set

  Validate on val_set

  Save best performing checkpoint (lowest val_loss)

end for

Step 5: Prediction and Classification

Use softmax on final vector to predict sentiment polarity:

  → Output ∈ {Positive, Negative, Neutral}

Pair aspect terms with sentiment predictions → Aspect-Sentiment Pairs

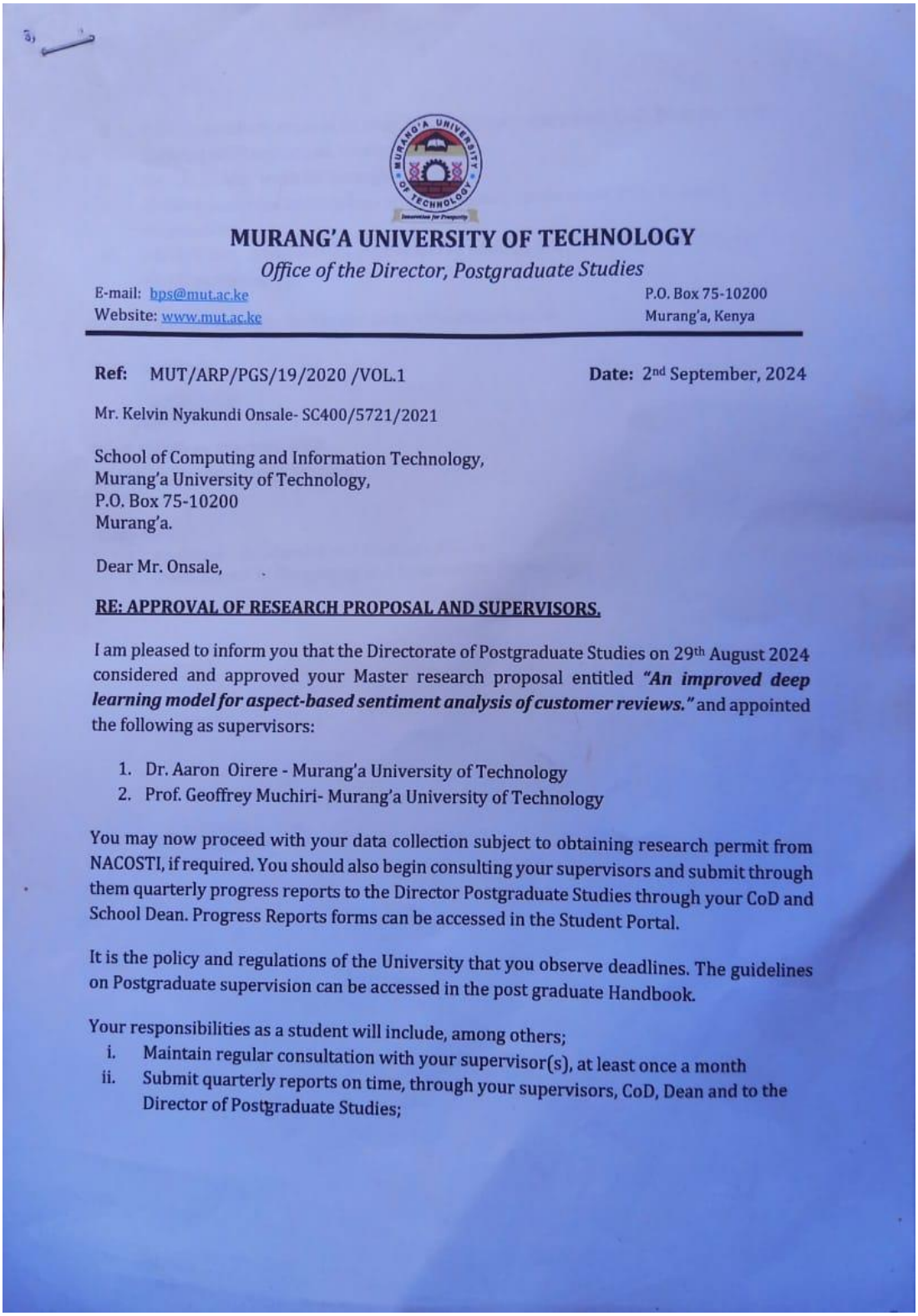
Step 6: Evaluation

Evaluate ***M*** on test_set using Accuracy, Precision, Recall, F1-Score

Return Trained_Model, Evaluation_Scores

End Algorithm

Appendix 4: Approval of Research Proposal and Supervisors



MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Director, Postgraduate Studies

E-mail: bps@mut.ac.ke
Website: www.mut.ac.ke

P.O. Box 75-10200
Murang'a, Kenya

Ref: MUT/ARP/PGS/19/2020 /VOL.1 **Date:** 2nd September, 2024

Mr. Kelvin Nyakundi Onsale- SC400/5721/2021

School of Computing and Information Technology,
Murang'a University of Technology,
P.O. Box 75-10200
Murang'a.

Dear Mr. Onsale,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 29th August 2024 considered and approved your Master research proposal entitled ***"An improved deep learning model for aspect-based sentiment analysis of customer reviews."*** and appointed the following as supervisors:

1. Dr. Aaron Oirere - Murang'a University of Technology
2. Prof. Geoffrey Muchiri- Murang'a University of Technology

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports forms can be accessed in the Student Portal.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.

Your responsibilities as a student will include, among others;

i. Maintain regular consultation with your supervisor(s), at least once a month
ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;

ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;
iii. Ensure quality work all through;
iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.
v. Publish one articles from your research findings in a refereed journal prior to thesis examination.

For any further clarification please contact the undersigned.

Yours Sincerely,

Prof. Geoffrey Muchiri, PhD
Director, Postgraduate Studies
GM/PIM

Copy to:
1. Registrar – Academics and Students Affairs
2. Dean, School of Computing and Information Technology

Appendix 5: NACOSTI Research License

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **929889**

Date of Issue: **13/September/2024**

RESEARCH LICENSE



This is to Certify that Mr.. Kevin Nyakundi of Murang'a University of Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Muranga on the topic: AN IMPROVED DEEP LEARNING MODEL FOR ASPECT-BASED SENTIMENT ANALYSIS OF CUSTOMER REVIEWS for the period ending : 13/September/2025.

License No: **NACOSTI/P/24/39888**

929889

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix 6: Publication

The following paper has been published from this thesis:

[1] K. N. Onsale, A. M. Oirere, and G. M. Muketha, "A Comparative Study of Attention Mechanisms in Deep Learning Models for Aspect-based Sentiment Analysis of Customer Reviews," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 14, no. 2, pp. 42–50, Apr. 2025, doi: https://doi.org/10.30534/ijatcse/2025/011422025.