A HYBRID DEEP LEARNING MODEL FOR DETECTING POLARITY SHIFT DUE TO IMPLICIT NEGATION IN SENTIMENT ANALYSIS OF MOBILE PHONE BRANDS' REVIEWS

Millicent Kathambi Murithi

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Information Technology of Murang'a University of Technology

July, 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 awarded in this or any other university. I understand what plagiarism is and I am aware of the university 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.

_____ _____

Millicent Kathambi Murithi Date

SC500/5532/2020

Supervisors' Approval

The undersigned certify that they have read and hereby recommend for acceptance of Murang'a University of Technology thesis entitled "**A Hybrid Deep Learning Model for Detecting Polarity Shift due to Implicit Negation in Sentiment Analysis of Mobile Phone Brands' Reviews".**

_____ _____

Dr. Aaron Mogeni Oirere, Date

Department of Computer Science,

Murang'a University of Technology.

_____ _____

Dr. Rachael Njeri Ndung'u, Date

Department of Information Technology,

Murang'a University of Technology.

DEDICATION

This thesis is dedicated to my husband Kinyua, my daughters Kendi and Makena and my son Gikera. Their moral support and encouragement in the entire process gave me perseverance to try my best.

ACKNOWLEDGEMENT

I am deeply grateful to God for his grace and strength throughout my PhD journey. This thesis would not have been possible without the unwavering support and guidance from a number of important people in my life.

First and foremost, I extend my heartfelt gratitude to my first supervisor Dr. Aaron Mogeni Oirere, from the Department of Computer Science at Murang'a University of Technology, Kenya. Your insightful suggestions and advice greatly enhanced the quality of my work. Am sincerely grateful to my second supervisor, Dr. Rachael Njeri Ndung'u from the department of Information Technology at Murang'a University of Technology, Kenya. Your constructive criticism and advices improved my work greatly. Thank you my supervisors for meticulously editing my manuscripts, offering numerous improvements that strengthened both the scientific rigor and academic style of my research. Your guidance was instrumental during the empirical setups and data analysis process.

I owe a deep debt of gratitude to my husband Kinyua and my three children—Kendi, Makena, and Gikera —whose constant encouragement and moral support motivated me to persevere through every challenge. Am grateful to my parents, brother, sisters and relatives for being a source of inspiration during my academic journey.

Lastly, I thank my classmates, colleagues and friends for always being there for me in times of consultation and always coming forth to offer insights.

ABSTRACT

Sentiment, the emotional tone, attitude, or opinion expressed in a piece of text, is critical as it captures the emotional polarity or subjective judgment embedded in a language. Sentiment analysis allows any brand or company to gain insight into how consumers feel about their products, brands, or the company itself through the variety of feedback that comes from the digital marketing platforms. These sentiments can have a positive, neutral or negative polarity. A problem arises with the existing sentiment analysis model when the sentiment of text is reversed which results in positive polarity being classified as negative polarity and the vice versa. This results into a polarity shift problem. Therefore, the classifier needs to be able to recognize polarity shifts and separate the impact that their scope has on the text's final polarity. Sentiment analysis models are evolving rapidly, with advancements ranging from state of the art data preprocessing techniques, machine learning, deep learning and transformers. Despite these advancements, the current models still face significant hurdles in accurately detecting polarity shift due to implicit negation, as their ability to handle nuanced shifts remains limited. Polarity shifts due to implicit negations present a persistent challenge in sentiment analysis, as they change the accurate interpretation of sentiment in written text. In light of this, this research is based on a pragmatism philosophy; firstly the current sentiment analysis models and the data preprocessing techniques are evaluated. This provides key findings and a conceptual framework that acts as the road map for the subsequent empirical analysis process. The empirical analysis process is guided by a comprehensive methodology based on mixed research methods for triangular validation. Experiments are conducted on the sentiment analysis models that demonstrate methodological rigor and novelty, in a variety of state of the art feature extraction, feature selection and data augmentation techniques. The invaluable insights based on the results' analysis of the experiments of the state of art model that included SVM, Naïve Bayes, LSTM and BERT models, evinces the significance of the data augmentation process as a critical leverage point in the effective detection of polarity by the superior deep learning models. This observation guided the implementation of a hybrid deep learning model, through a multi-methodological model development methodology. The novelty of the proposed model lies in its comprehensive integration of Seq2Seq perturbations and contextual back translation for data augmentation, combined with FastText and BERT embeddings, and attention mechanisms, enabling it to effectively address polarity shifts due to implicit negation in sentiment analysis, particularly in mobile phone brand reviews. This model was then validated against the existing state of the art models, using similar metrics, in order to evaluate its effectiveness. Statistical significance tests, using the ANOVA and the Kruskal-Wallis H statistic, demonstrated that the new model is superior. This was also evinced by the improved average performance scores, attention heat maps as well as the presence of shorter whiskers and higher median (Q2) values in the whisker-box plots, in a variety of mobile phone reviews datasets. Future work could explore on combining the seq2seq perturbations on back translation with multiple data augmentation techniques to further enhance the model's ability to detect implicit negation and other nuanced shifts in sentiment. Moreover, future research could also explore methods for optimizing the training and inference efficiency of the hybrid model without sacrificing its efficacy. Techniques such as knowledge distillation, pruning, or quantization could help make the hybrid model more computationally efficient, enabling real-time sentiment analysis in resource-constrained environments.

TABLE OF CONTENTS

DECLARATION ..... **ii**
DEDICATION ..... **iii**
ACKNOWLEDGEMENT ..... **iv**
ABSTRACT ..... **v**
LIST OF TABLES ..... **xii**
LIST OF FIGURES ..... **xiii**
ACRONYMS AND ABBREVIATIONS ..... **xv**
CHAPTER ONE: INTRODUCTION **1**
1.1 Background of the Study 1
1.2 Problem Statement 5
1.3 Objectives 7
1.3.1 General Objective 7
1.3.2 Specific Objectives 8
1.4 Research Questions 8
1.5 Research Hypothesis 8
1.6 Significance of the study 9
1.7 Scope of the study 10
1.8 Limitations of the Study 11
1.9 Contributions of the Thesis 11
1.9.1 Contributions to Theory 11
1.9.2 Contributions to Practice 13
1.10 Organization of the Thesis 13
CHAPTER TWO: LITERATURE REVIEW **15**
2.1 Introduction 15
2.2 Theoretical Foundations of Sentiment Analysis 16
2.2.1 Psychology Theory 16
2.2.2 NLP Theory 17
2.2.2.1 Cognitive Linguistic Theory 17
2.2.2.2 Text Representation Theory 18
2.2.2.3 NLP Techniques 19
2.2.3 Machine Learning Theory 19
2.2.3.1 Supervised Learning Theory 19
2.2.3.2 Representation learning 20
2.2.3.3 Optimization Theory 20
2.2.3.4 Regularization Theory 20
2.2.3.5 Hybrid Learning Theory 20
2.2.4 Information Retrieval and Text Mining Theory 21

2.3 Architecture of Sentiment Analysis Models 21
2.3.1 Data Collection and Standardization 23
2.3.2 Data Pre-processing in Sentiment Analysis 23
2.3.2.1 Feature Selection 26
2.3.2.2 Data Augmentation Preprocessing 28
2.3.3 Feature Extraction Process in Sentiment Analysis 29
2.3.3.1 TF-IDF (Term Frequency and Inverse Document Frequency) Approach 30
2.3.3.2 Use of Bag-of-Words 32
2.3.3.3 Word Embedding 32
2.3.3.4 Count Vectorizer 34
2.3.3.5 N-grams 34
2.3.4 Sentiment Classification Process 35
2.3.5 Sentiment Summarization 37
2.4 Factors Affecting Polarity Shifts in Sentiment Classification 38
2.5 Review of the Existing Sentiment Analysis Models for Handling Polarity Shift 42
2.5.1 Sentiment Analysis Models used in Detecting and Handling the Polarity Shift Problem 44
2.5.2 Polarity Shifters Commonly Detected by the Existing Sentiment Analysis Models 48
2.5.3 The Input Datasets used with the Sentiment Analysis Models for Detecting Polarity Shift Problem 50
2.5.4 Feature Selection Techniques applied in the Sentiment Analysis Models used for Detecting and Handling the Polarity Shift Problem 51
2.5.5 Metrics used to assess the effectiveness of the Sentiment Classification Models used for Detecting the Polarity Shifts 52
2.5.6 Discussions on the Literature Review Analysis 54
2.5.6.1 Sentiments Classification Models' Efficacy and Future Research 54
2.5.6.2 Common Polarity Shifters and the Dynamism in Handling Polarity Shift 54
2.5.6.3 Trends in Feature Selection and Feature Extraction Techniques used with Sentiment Classification Models 55
2.5.6.4 Domain-Specific Considerations for the Datasets and the Labelling Process 56
2.5.6.5 Performance Evaluation Metrics 56

2.6 Implications of the Key Findings in the Experimental Design Process 57
2.7 Conceptual Framework 58
2.7.1 Independent Variables 59
2.7.2 Moderating Variables 59
2.7.3 Dependent Variables 59
2.8 Summary 60
CHAPTER THREE: METHODOLOGY 63
3.1 Introduction 63
3.2 Research Philosophy 63
3.3 Research Design 64
3.3.1 Systematic Literature Review Methodology 66
3.3.2 Pilot Study 67
3.3.3 Experimental Design for the Empirical Baseline Analysis 67
3.3.3.1 Main Experiments 68
3.3.3.2 Population Sampling Strategy and Sample Size 70
3.3.3.3 Research Instruments 79
3.3.3.4 Integrated Development Environment and Libraries 80
3.4 Experimental Data Collection and Techniques 83
3.5 Experimental Data Analysis 85
3.6 Research Ethics 87
3.7 Summary 88
CHAPTER FOUR: EVALUATION OF THE EXISTING SENTIMENT ANALYSIS MODELS FOR DETECTING POLARITY SHIFT 89
4.1 Introduction 89
4.2 Pilot Study Results 89
4.3 Experimental Results 90
4.3.1 One-Way ANOVA Analysis of the Average Normalized Score on Sentiment Classification Model 93
4.3.2 Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Extraction Techniques 95
4.3.3 Two-way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Selection Techniques 101
4.3.4 Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Data Augmentation 110
4.3.5 Three-Way ANOVA Analysis of the Average Normalized Score for Sentiment Analysis Model, Feature Extraction Method and Feature Selection Techniques 112

4.3.6 Three-Way ANOVA Analysis of the Average Normalized Score for Sentiment Analysis Model, Feature Selection Method and Data Augmentation Method 113
4.3.7 Three-Way ANOVA Analysis of the Average Normalized Score for Feature Extraction Method, Feature Selection Method and Data Augmentation Method..... 115
4.3.8 Three-way ANOVA analysis of the average normalized score for Sentiment Analysis Model, Feature Extraction Method and Data Augmentation technique..... 116
4.3.9 Bar Graph of the Total Average Normalized Scores of the Different Feature Extraction Methods 117
4.3.10 Bar Graph of the Total Average Normalized Scores of the Different Feature Selection Methods..... 118
4.3.11 Bar Graph of the Total Average Normalized Scores of the Different Data Augmentation Techniques..... 119
4.3.12 Bar Graph of the Total Average Normalized Scores of the Different Categories of Sentiment Analysis Models 120
4.4 Discussion 121
4.4.1 Significant Differences between Traditional ML and Deep Learning Models..... 121
4.4.2 Interactions between Sentiment Analysis Models and Feature Extraction Techniques 122
4.4.3 Feature Selection Techniques Impact on Sentiment Analysis..... 122
4.4.4 Data Augmentation Techniques and its Impact in Sentiment Analysis Model..... 123
4.5 Key Findings from Evaluation Analysis..... 125
4.6 Summary..... 126
CHAPTER FIVE: HYBRID DEEP LEARNING MODEL DEVELOPMENT 127
5.1 Introduction..... 127
5.2 Model Development Methodology 127
5.3 Model Architecture Design..... 128
5.4 Model Development..... 129
5.4.1 Data Preprocessing of the Proposed Hybrid Model..... 130
5.4.2 Model Architecture of the Proposed Hybrid Deep Learning Model..... 132
5.4.3 Model Training 134
5.4.4 Model's Algorithm..... 139
5.5 Discussion..... 142

5.6 Summary 143
CHAPTER SIX: VALIDATION AND COMPARATIVE EVALUATION OF THE HYBRID MODEL 144
6.1 Introduction 144
6.2 Hybrid Model Validation Approach 144
6.3 Hybrid Model Validation Tests Results 146
6.3.1 Average Performance Score of the Newly Developed Sentiment Analysis Model against the Existing Ones 146
6.3.2 The ANOVA Analysis of the Average Normalized Score of the Newly Developed 148
6.3.3 Kruskal Wallis H Statistic of the Newly Developed Sentiment Analysis Model against the Existing Ones 149
6.3.4 Mean Squared Error of the Newly Developed Sentiment Analysis Model against the Existing Ones 151
6.3.5 Whisker Plots of the Newly Developed Sentiment Analysis Model against the Existing Ones 152
6.3.6 Attention Heat Maps of the Newly Developed Sentiment Analysis Model against the Existing Ones. 153
6.3.7 Hybrid Model Performance Visualization using SHAP 155
6.4 Model Validation Triangulation 158
6.5 Error Analysis of the Hybrid Deep Learning Model 159
6.6 Discussion 162
6.7 Hybrid Model's Limitations and Future Improvements 163
6.8 Summary 165
CHAPTER SEVEN: SUMMARY, CONCLUSION AND RECOMMENDATIONS 166
7.1 Summary 166
7.2 Conclusion 168
7.3 Recommendation on Policy and Practice 171
7.3.1 Integration with Business and Decision Systems 171
7.3.2 Testing with Contextual Embedding for Implicit Negation 171
7.3.3 Testing with Hybrid Model Design 171
7.3.4 Adoption of Comprehensive Evaluation Metrics 172
7.4 Recommendations for Future Research 172
7.4.1 Experiments with other Training Optimization Approaches 172
7.4.2 Experiments with Computation Resource Efficiency Environment 173
7.4.3 Experiment with Multimodal Sentiment Analysis 173

7.4.4 Experiment with Mixed Sentiments and Contextual Contradictions 173
REFERENCES 174
APPENDICES 199

LIST OF TABLES

Table 3.1: Libraries in Different Model Categories 81
Table 4.1: Summary of the Set of Variables, Interactions and the Data Analysis Method Applied in the Different Set of Experiments. 92
Table 4.2: One-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model 93
Table 4.3: Post Hoc Analysis of the Average Normalized Score on Sentiment Analysis Model 94
Table 4.4:Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Extraction Techniques 96
Table 4.5: Post Hoc Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Extraction Techniques 96
Table 4.6: Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Selection Techniques 102
Table 4.7: Post Hoc Analysis of Sentiment Analysis Models and Feature Selection Techniques 102
Table 4.8:Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Data Augmentation 111
Table 4.9: Post Hoc Analysis of Model and Data augmentation techniques 111
Table 4.10: Three-Way ANOVA of the Sentiment Analysis Models, Feature Extraction Techniques and the Feature Selection Techniques 113
Table 4.11: Three-Way ANOVA of the Sentiment Analysis Models, Feature Selection Techniques and the Data Augmentation Techniques 114
Table 4.12: Three-Way ANOVA of the Feature Extraction Techniques, Feature Selection Techniques and the Data Augmentation Techniques 115
Table 4.13: Three-Way ANOVA of the Sentiment Analysis Models, Feature Extraction Techniques and the Data Augmentation Techniques 116
Table 6.1: ANOVA of the APS of the Developed Hybrid Model 148
Table 6.2: PostHoc analysis of the APS of the developed Hybrid model 148
Table 6.3 Kruskal-wallis Results of the APS of the Hybrid Model 150
Table 6.4 Dwass -Steel- Critchlow-Fligner Pairwise Comparisons 150
Table 6.5: Qualitative Error Analysis of the Hybrid Model Based on a Sample of Amazon Mobile Phone Reviews 159

LIST OF FIGURES

Figure 1.1: Taxonomy of sentiment analysis 2
Figure 1.2 : Explicit and implicit features of the mobile phone reviews 3
Figure 1.3: Taxonomy of sentiment analysis techniques [17] 4
Figure 2.1: Sentiment Analysis Framework 22
Figure 2.2: Flow Chart of the Data Cleaning and Preprocessing [48] 24
Figure 2.3: Lemmatization of the Mobile Phone Reviews [57] 26
Figure 2.4: Tokenization Process of the Mobile Phone Reviews [47] 26
Figure 2.5: Sentiment Classification Techniques [5] 36
Figure 2.6: PRISMA Flow Chart for the Articles Selection[111] 43
Figure 2.7: Research Trends on the Sentiment Analysis Models [124] 48
Figure 2.8: Relationship between independent, moderating and dependent variables 60
Figure 3.1: Mapping of research objectives, and questions to research methods 66
Figure 3.2: Activity Diagram for the Empirical Analysis 78
Figure 4.1: Bar graph of the total normalized average scores for feature extraction 117
Figure 4.2: Bar graph of the total normalized average scores for the different feature selection techniques 118
Figure 4.3: Bar graph of the total normalized average scores for the different data augmentation techniques 119
Figure 4.4: Bar graph of normalized performance scores of sentiment analysis models 120
Figure 5.1: Model Development Methodology 128
Figure 5.2: Data Preprocessing of the Hybrid Model 131
Figure 5.3 Proposed Hybrid Model Architecture 133
Figure 5.4: Model Training, and Testing Framework. 135
Figure 5.5: Loss of the Hybrid Deep Learning Model for Detecting Polarity Shift 136
Figure 5.6: APS of the Hybrid Deep Learning Model for detecting polarity shift 138
Figure 6.1 Model Validation Stages 145
Figure 6.2: Average Performance Score of the Newly Developed Sentiment Analysis Model against the Existing Ones 147
Figure 6.3: Mean Squared Error of the Newly Developed Model against the Existing Ones 151
Figure 6.4: Whisker Plots of the Newly Developed Sentiment Analysis Model against the Existing Ones 153

Figure 6.5: Attention Heat Maps of the Newly Developed Sentiment Analysis Model against the Existing Ones 154
Figure 6.6: Sentiment Score Distribution of the Combined Title and Body Features in Amazon Mobile Phone Datasets 156
Figure 6.7: Sentiment Score Distribution of Title and Body Features in Amazon Mobile Phone Reviews Datasets 157

ACRONYMS AND ABBREVIATIONS

ANOVA:	Analysis of Variance
APS:	Average Performance Score
BoW:	Bag of Words
CBoW:	Continuous Bag of Words
CNN:	Convolutional Neural Network
CRF:	Conditional Random Field
EDA:	Exploratory Data Analysis
GPU:	Graphics Processing Unit
HMM:	Hidden Markov Model
IDE:	Integrated Development Environment
IEEE:	Institute of Electrical and Electronics Engineers
LSTM:	Long Short-Term Memory
MLP:	Multi-Layer Perceptron
NACOSTI:	National Commission for Science, Technology and Innovation
NLP:	Natural Language Processing
NLTK:	Natural Language Processing Toolkit
PCA:	Principal Component Analysis
PSO:	Particle Swarm Optimization
RFE:	Recursive Feature Elimination
RNN:	Recurrent Neural Network
ReLU:	Rectified Linear Unit
SA:	Sentiment Analysis
SENTI-TFIDF:	Sentiment Term Frequency Inverse Document Frequency
TF-IDF:	Term Frequency Inverse Document Frequency
VSM:	Vector Space Model

CHAPTER ONE

INTRODUCTION

1.1 Background of the Study

The proliferation of the social media platforms has had a significant effect on how people express their opinions about products, companies, and services. The explosive expansion of the internet has made vast amounts of data become available, offering rich opportunities for market research and social media analysis [1]. These online activities generate a mix of structured and unstructured data characterized by high volume, velocity, variety, value, and variability. Effectively utilizing this data requires advanced natural language processing (NLP) approaches [1]. Among these, sentiment analysis has emerged as a pivotal tool for identifying the emotional undertone in text [2].

Sentiment analysis, a key aspect of NLP, involves extracting and classifying opinions expressed in text into positive, negative, or neutral polarities [3]. It operates at various levels i.e. document, sentence, and aspect levels, with each offering unique insights. At the document level, the overall sentiment of a document is assessed [4]. The sentence level evaluates polarity without isolating individual features, determining whether a sentence expresses a positive or negative sentiment. Aspect-level analysis focuses on specific features (aspects) of an entity, distinguishing between explicit and implicit features [5]. Explicit features are directly mentioned in the text, while implicit features, though not explicitly stated, are inferred from context. For example, "The Samsung phone has pretty good features" conveys explicit positive sentiment, whereas "Samsung phones are designed with a variety of features that cater to diverse user needs, often encouraging users to spare more time to fully adapt to the experience". This implicitly conveys negative sentiment that Samsung phones have a steep learning curve or requires great effort for users to fully utilize their features. Pinpointing this negative sentiment directly is a difficult task as it is very implicit besides being framed in a positive tone. According to [6], approximately 30% of reviews lack explicit opinion words but still exhibit clear sentiment orientation, termed implicit sentiment.

Figure 1.1 illustrates sentiment analysis as a classification into document, sentence, and aspect levels, with the latter comprising both the implicit and the explicit aspects.

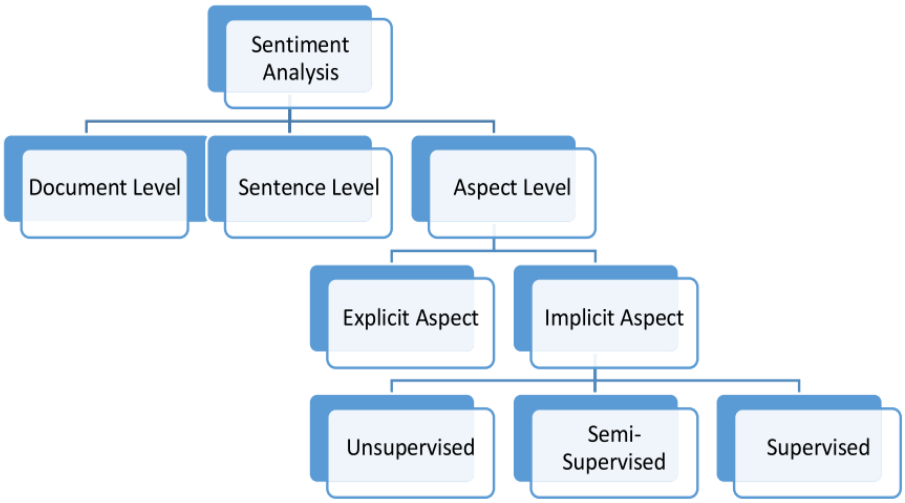


Figure 1.1: Taxonomy of sentiment analysis [6]

In the last decade, advancements in telecommunications technology have made mobile phones indispensable. The diverse features on varied user preferences make consumers face challenges in selecting suitable devices. User-generated reviews provide insights into specific features and overall performance, aiding decision-making. Sentiment analysis enhances this process by extracting and analysing consumer opinions about mobile phone features, helping to resolve polarity shifts and improve classifier performance [7]. An example of both explicit and implicit features for mobile phone reviews based on the given comments is shown in Figure 1.2. The customers' reviews contain explicit aspects with also an important portion of the opinion aspect that is not mentioned explicitly but implied implicitly [8].

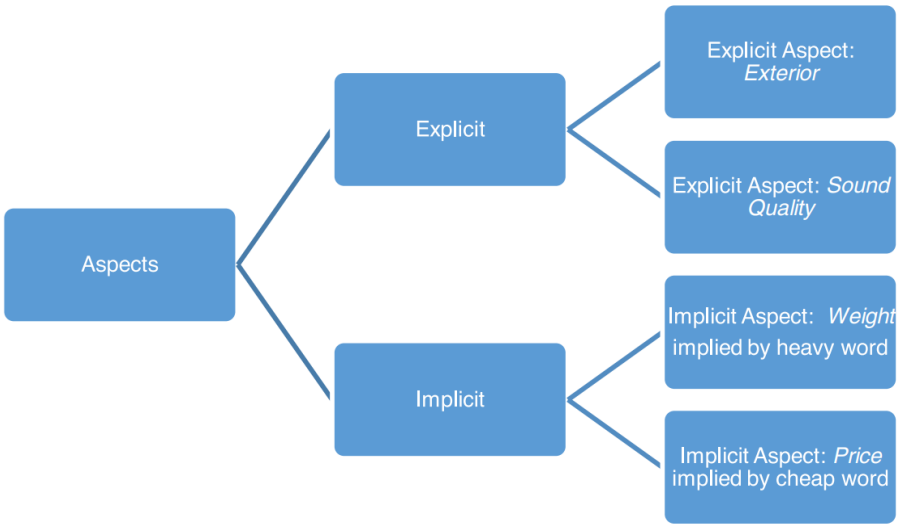


Figure 1.2 : Explicit and implicit features of the mobile phone reviews [8]

The task of understanding and interpreting sentiment in text is both challenging and essential [9]. Traditional sentiment analysis approaches employ data mining, machine learning, artificial intelligence, and computational linguistics [10]. Recent advancements in NLP have introduced sophisticated models incorporating deep learning, pre-trained embeddings, and attention mechanisms [11]. However, polarity shifts where positive polarity reverses to negative or vice versa remains a significant challenge. Polarity shifts, often caused by negation, contrast, or sentiment inconsistencies, can dramatically alter sentence meaning [12] leading to inaccurate sentiment classification. This leads to business making wrong decisions on their product and services. Detecting polarity shifts accurately is critical for enhancing sentiment analysis accuracy. While state-of-the-art models claim contextual and semantic understanding, their ability to handle sentiment flux caused by implicit negation requires further investigation [13]**;**[14]**.** Implicit negation, where sentiment contradicts the literal meaning of words, introduces additional complexity [15]. As language evolves, accurately detecting these subtleties becomes increasingly important for applications like customer feedback analysis and social media monitoring [16].Sentiment analysis approaches fall into three main categories i.e. lexicon-based, machine learning-based, and hybrid methods [16]. Figure 1.3 presents the taxonomy of sentiment analysis techniques.

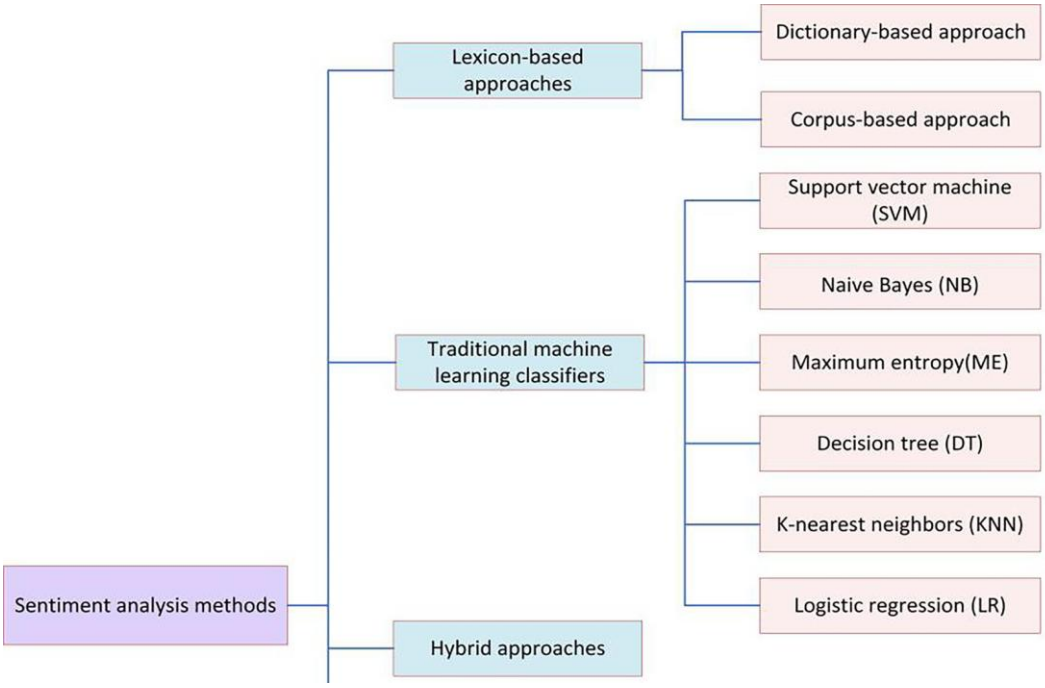


Figure 1.3: Taxonomy of sentiment analysis techniques [17]

Lexicon-based methods rely on emotional lexicons and are context and language-dependent. Since machine learning approaches, treats sentiment analysis as a text classification problem, they employ classifiers like Naive Bayes, Support Vector Machines (SVM), and logistic regression. Both supervised and unsupervised learning techniques have been applied to sentiment classification [17]. Hybrid methods combine the strengths of both approaches, offering improved flexibility and robustness.

Recently, sentiment analysis has seen advancement in deep learning, natural language understanding, and hybrid approaches. These trends emphasize not only improved accuracy but also greater adaptability, domain sensitivity, and explainability. Some of these advancements include transformer based architectures with pre-trained embedding models such as BERT and RoBERTa [18]. Other advancement includes multilingual and cross-lingual sentiment analysis that can handle diverse languages like the Multilingual BERT (mBERT). Integration of tools like SHAP, LIME, and Integrated Gradients enables the interpretability of the sentiment classification models [19]. Due to increasingly application of sentiment analysis in sensitive areas like finance and healthcare, explainability has become a key concern. Attention visualization has been adopted in transformers to provide

insights into which words or phrases influenced sentiment classification. Furthermore, data augmentation has emerged as a key strategy in sentiment analysis, especially to combat data sparsity, improve generalization, and reduce model overfitting. Recent advancements in augmentation techniques go far beyond traditional synonym replacement and back translation, leveraging deep learning and NLP-specific innovations. Data augmentation has seen the evolution of contextual embedding-based augmentation and use of pre-trained BERT to predict replacements. Seq2seq perturbation has been used in sentiment analysis to introduce syntactic or lexical variation without altering meaning[20]. More specifically seq2seq has been used as sentiment-preserving paraphrasing especially with data augmentation techniques. For example it generates multiple reworded versions of the same input that carry the same sentiment label.

1.2 Problem Statement

The extraction of subjective information such as emotions, attitudes, and opinions from textual data, through sentiment analysis, typically fall under three polarities categories i.e. positive, negative, and neutral. Positive polarity sentiments indicate favor towards a product, neutral polarity represents text that lacks strong emotional cues, making it objective while negative polarity sentiments express criticism [14]. Accurate sentiment classification is critical for businesses and consumers, for shaping product success and customer engagement [9]. However, one of the central challenges in sentiment analysis is the inaccurate detection of polarity shift, where the sentiment of a statement is unexpectedly detected as positive instead of negative or vice versa [22].

Such shifts are triggered by various factors, including explicit and implicit negation, contrast, ambiguity, and evolving linguistic patterns [23];[24]. However, detecting polarity shifts remains a significant challenge, particularly when implicit negation alters the intended sentiment of a statement [5]. An advanced pre-processing technique, such as standard normalization that takes deliberately into account the case of negation and mixed emotion would be extremely beneficial. Therefore, the usage of sentiment dictionaries with pre-defined polarities for words is not an appropriate practice for sentiment analysis. Existing sentiment dictionaries fail to handle sarcasm or negations to a great extent [21]. It is observed that many machine learning based techniques are trained to work only for a particular domain. They do

not consider that the words can change their polarity based on the context and domain of application. Thus, whenever the same word is tested for another domain using the trained classifier, it shows incorrect results in some situations.

Recent advancements in deep learning models, including the use of pre-trained embeddings and attention mechanisms have led to state-of-the-art models capable of understanding contextual nuances and semantic complexities [11]. Despite these strides, the ability of these models to effectively and accurately detect polarity shift due to implicit negation due to context and domain application remains an area that demands rigorous investigation in the field of sentiment analysis and natural language processing in general [13].

Existing deep learning models lack fine-grained sentiment parsing whereby they struggle to distinguish between different parts of the text that carry different sentiments leading to an incorrect or generalized sentiment prediction for the entire text [22]. Implicit negation often relies on contextual cues and subtle shifts in meaning, which are not as directly observable as explicit negation markers such as "not" or "never". For example, the phrase "The phone could have been better" implies dissatisfaction, even though no direct negation is present. Existing sentiment analysis models struggle with this issue because they primarily rely on explicit negation markers (e.g., ‘not,’ ‘never’), failing to recognize context-dependent sentiment reversals. This limitation affects classification accuracy, leading to misinterpretations in consumer feedback, brand reputation analysis, and automated decision-making systems [23]. Addressing this challenge requires a model capable of detecting polarity shifts beyond explicit linguistic cues, ensuring more precise sentiment classification in real-world applications.

The problem of implicit negation and polarity shift has been previously addressed by a number of researchers. For instance, Li et al. [6] introduced a supervised contrastive learning model that captures both explicit and implicit sentiment shifts by aligning sentiment representations in an embedding space. However, this approach predominantly focuses on explicit polarity markers and is limited in handling the broader range of implicit negations that occur in real-world data. Similarly, Syafiqah and Nafis [24] proposed the Senti-TFIDF model, which effectively handles explicit negation but fails to address the complexities of implicit negation or hidden

sentiment shifts. Other models, such as those utilizing long short-term memory (LSTM) [25], have made strides in recognizing polarity shifts, but they often focus on a narrow subset of shift types and still struggle with capturing implicit negation's full range of impact [23]. These existing models primarily tackle explicit negation with a limited exploration on the implicit negation. They focus on rule-based approaches, which are limited in their scope for handling implicit negation, and require significant manual effort to develop comprehensive sets of rules. Furthermore, these methods often lack the robustness needed to handle the diversity of polarity shifters encountered in natural language, such as contrast, contextual ambiguity, and cultural nuances [26].

While recent research [27] has suggested that combining various factors such as contrast transitions, sentiment inconsistency, and implicit negations could enhance sentiment classification, current methods still fail to offer a scalable and highly accurate solution.

Despite enhancement of sentiment analysis models accuracy through incorporation of feature extraction, selection, and data augmentation techniques, the problem of polarity shift due to implicit negation remains an intractable and open research challenge. This research, therefore, aims to develop a hybrid model that effectively addresses polarity shifts caused by implicit negation, providing enhanced accuracy and reliability in sentiment classification. This novel model integrates both the theoretical insights from the literature analysis and the empirical insights from a detailed ANOVA analysis of experimental results, ensuring a theoretically grounded and methodologically sound approach to the problem solution. The novel model is a hybrid of a transformer-based BERT and LSTM deep learning models coupled with a contextual back translation augmented by Seq2seq perturbations tailored for implicit negation detection.

1.3 Objectives

1.3.1 General Objective

This research aims to develop a hybrid deep learning model for detecting polarity shifts due to implicit negation in sentiment analysis of the mobile phone brands' reviews based on insights of both the theoretical and empirical analysis.

1.3.2 Specific Objectives

The specific research objectives of this research were formulated as follows:

1. To investigate the implicit sentiment inversion factors that affect polarity shifts in sentiment classification with the aim of identifying the critical polarity shifters in written texts.
2. To evaluate the existing sentiment analysis models used in detecting polarity shifts with the aim of comparing the performance among the conventional and hybrid models.
3. To develop deep hybrid model for detecting the polarity shift due to implicit negation of the mobile phone brand reviews in order to improve the effectiveness of sentiment classification.
4. To perform a comparative evaluation of the developed model against the existing ones using multiple metrics and datasets**.**

1.4 Research Questions

The specific research questions are formulated as follows:

1. What implicit sentiment inversion factors affect polarity shifts in sentiment classification within written texts?
2. How do the performance among the conventional and hybrid models of the existing sentiment classification models compare in the detection of polarity shifts?
3. How can an effective deep hybrid sentiment classification model for detecting the polarity shift due to implicit negation be developed, based on the results of the comparative analysis process?
4. How can the developed model be evaluated through a comparative analysis against the existing models using multiple metrics and datasets?

1.5 Research Hypothesis

In this research study, both the null hypothesis (H0) and the alternate hypothesis (Ha) were formulated as follows:

H_{a0}: There are no statistically significant differences among the performance of the sentiment analysis models used for detecting polarity shift due to implicit negation.

H_{a1}: There are statistically significant differences among the performance of the sentiment analysis models used for detecting polarity shift due to implicit negation.

1.6 Significance of the study

The significance of this research is both theoretical and practical, addressing key challenges in sentiment analysis while providing actionable insights for real-world applications. Theoretically, this research makes a substantial contribution to the field of natural language processing and sentiment analysis by advancing the understanding of polarity shifts triggered by implicit negation. While polarity shifts have been recognized as a challenge in sentiment analysis, the role of implicit negation in altering sentiment interpretation remains inadequately addressed. This study offers a comprehensive analysis of how implicit negations, often subtle and context-dependent, affect sentiment classification, adding depth to existing knowledge about language complexities and sentiment interpretation. By identifying critical factors that influence polarity shifts, this research expands the theoretical framework of sentiment analysis models, providing new avenues for refining these models to handle more nuanced language expressions effectively. This, in turn, enhances the theoretical understanding of sentiment analysis in both computational linguistics and machine learning. Practically, the development of a novel model to accurately detect and manage implicit negation in sentiment analysis has significant implications across various industries. This model represents an innovation in the field, with direct applications in customer feedback analysis across sectors such as finance, hospitality, and healthcare. In these domains, accurate sentiment interpretation is essential for making informed decisions based on customer opinions, reviews, and interactions. Furthermore, this research holds implications for brand monitoring, social media sentiment analysis, and market intelligence, where companies rely on sentiment data to shape strategies, understand consumer preferences, and improve services. The enhanced model's ability to handle implicit negation ensures more accurate sentiment classification, leading to more reliable insights in these critical areas.

Moreover, the validation of this novel model, particularly its superior performance in comparison to existing approaches, will establish it as a benchmark for future

research and development in sentiment analysis. By addressing the limitations of current models, this work encourages the adoption of more sophisticated methods that can better manage complex linguistic phenomena.

This study not only bridges a critical gap in sentiment analysis but also sets the stage for further advancements in the field, promoting the evolution of models that are better equipped to interpret subtle, context-dependent shifts in sentiment.

1.7 Scope of the study

The scope of this research focuses on the sentiment analysis models for detecting polarity shifts due to implicit negations. The study first explores the polarity shifters and the efficacy of sentiment analysis models, used in detecting the polarity shift due to implicit negation. The main focus was the negation polarity shifter as evidenced by the existing literature that it is the most prevalent polarity shifter in sentiment analysis. This is achieved through a comprehensive literature review analysis that forms the foundation for understanding implicit negation and its effects on polarity shift in sentiment analysis. The key findings from the literature review, theoretical frameworks and linguistic constructs influencing implicit negation are incorporated in the subsequent experimental research design in the empirical analysis process. Secondly, an empirical analysis is conducted using the amazon mobile phone reviews. The amazon mobile phone reviews is appropriate for this study because it is widely regarded as a benchmark in sentiment analysis research due to its large scale, diversity in language, and the presence of complex sentiments that may shift due to implicit negation. Moreso, its availability and extensive metadata, such as product attributes and user demographics, further enrich its utility for model training. The empirical analysis for the baseline models is implemented using the jupyter notebook on the google colab IDE including a set of python libraries for the state of the art models, feature extraction, feature selection, and data augmentation techniques. The choice of the techniques used in the empirical study was guided by a preliminary pilot study. In order to validate the findings and generate insights from the empirical data, the ANOVA statistical technique is applied and culminates to a set of key empirical findings. Thirdly, a novel hybrid deep learning model for detecting polarity shift due to implicit negation is developed, integrating the insights from the

empirical analysis and underpinned on the principles of the theoretical literature review analysis.

1.8 Limitations of the Study

This section highlights the following limitations encountered during this research study:

i. During reviewing of the literature, there was multi domain set of datasets and varying evaluation metrics across the various models. This made it difficult to compare the performance of the different sentiment classification models.

   To overcome this limitation, a Single domain datasets and consistence evaluation metric was adopted during the pilot study and the empirical study to ensure fairness during the evaluation of the different models.

ii. During the empirical analysis process, different sentiment classification models converges over varying numbers of runs and iterations, this made it difficult to report uniform stopping criteria values across all models. Some models require more iterations than others within a single run before achieving convergence, while others need more runs before stability is reached.

   This limitation was addressed by setting stopping criteria for both runs and iterations based on the maximum values exhibited by the models that require the highest thresholds to converge. Consequently, the maximum iterations for all techniques are uniformly set to 10, even in cases where some models converge in less iteration. Additionally, the maximum number of runs for all models is set to a uniform value of 15.

1.9 Contributions of the Thesis

This research study enhances the researcher's knowledge base, primarily grounded in the classification theory of machine learning, in the following ways:

1.9.1 Contributions to Theory

i. The factors that affect the polarity shift in sentiment classification models are identified through reviewing of the literature. According to the review,

negation emerges as the most significant polarity shifter. It is notable that polarity shift due to implicit negation is a research area that requires more work. This research focused on the development of contextual back translation augmented by Seq2seq perturbations with the aim of detecting the implicit negation.

ii. The empirical analysis process provides invaluable insights into both practical and theoretical research contributions of this thesis. The theoretical implications of our findings extend beyond model development by offering a deeper understanding of the interplay between different techniques in the context of sentiment analysis on detecting implicit negation. These techniques include feature selection, feature extraction and data augmentation used in the sentiment analysis. The empirical analysis review that there is an interaction between feature selection and feature extraction with the traditional sentiment classification model but not with the deep learning models. At the same time, there is interaction between data augmentation techniques and deep learning models used in sentiment classification and not with the traditional sentiment classification models.

   This research study adopted a hybrid of deep learning (BERT and LSTM) and an enhanced data augmentation through contextual back translation inorder to capture nuanced polarity shifter in varying context. The practical implication will lead to a more accurate sentiment classification of the mobile phone reviews.

iii. The improvement of the sentiment data augmentation techniques through the integration of contextual back translation and Seq2seq perturbations leads to improved implicit negation detection. This allowed for the generation of a diverse interpretation that consequently improves the model's ability to handle nuanced expressions of sentiments due to implicit negation.

   The development of deep hybrid model with an improved data augmentation process through seq2seq perturbations improves sentiment classification accuracy, generalizability, robustness, and contextual understanding in mobile phones reviews. This approach addresses the identified challenges of detecting polarity shift due to implicit negations.

iv. The proposed model's performance demonstrated statistical significance difference compared with the four best performing sentiment classification models identified during empirical analysis. The models' performance was tested using three mobile phone datasets publically available. These include: amazon mobile reviews, google play store reviews and iPhone App store reviews.

The comparative analysis of feature selection, feature extraction, data augmentation and sentiment classification techniques, identifies optimal strategies for context-dependent polarity shifts. This allowed for cross platform comparison that helps assess the model's robustness across distinct operating system and user expectations.

1.9.2 Contributions to Practice

i. The hybrid deep learning model can be adopted across different domains to accurately classify sentiments through improved detection of implicit negation polarity shifters. This will help organization make informed decision based on the accurate classification of users' sentiments about a product or a service.
ii. The hybrid model helps with the brand monitoring, social media sentiment analysis, and market intelligence, where companies rely on sentiment data to shape strategies, understand consumer preferences, and improve services

1.10 Organization of the Thesis

The subsequent sections in this thesis are organized as follows: The theoretical foundations of sentiment analysis, an assessment of the literature on sentiment analysis polarity shifters, and models for handling polarity shift is presented in Chapter two with conceptual framework and the summary of the findings being the culmination.

Chapter three, the research methodology provides details on how this research was undertaken. It entails the research philosophy, the research design, experimental data collection techniques, and data analysis and research ethics.

Chapter four describes the empirical insights into the state of the art sentiment analysis models used in handling the polarity shift. This culminates to the one-way, two-way and three-way ANOVA analysis of the experimental data and finally the key findings from the empirical analysis are discussed.

Chapter five presents the hybrid deep learning model development along with its implementation process. Model development methodology, model architecture design and development process is discussed.

Chapter six presents the hybrid model validation and comparative evaluation results based on a number of datasets and the baseline models. Lastly, chapter seven summarizes and concludes the thesis as well as highlights recommendations for future work in the area of study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

The literature review focuses on a comprehensive exploration of the sentiment analysis in polarity shift detection. Firstly, the fundamentals of sentiment analysis models including their architecture as well as data preprocessing techniques are discussed. Each model's strategy, based on the sentiment analysis theory, is then evaluated for its effectiveness in detecting polarity shifts under the constraints and challenges inherent in implicit negation. The comparative study, based on a theoretical analysis process, and following the PRISMA approach, presents a number of the sentiment analysis models for handling polarity shifts, offering insights into their relative efficacy and limitations. Feature extractions, feature selection as well as the data augmentation techniques are explored as crucial data preprocessing tools during the process of polarity shift detection. The literature also delves into the metrics used to evaluate the effectiveness of these sentiment analysis models. The discussion on these metrics is done in the context of their relevance and reliability in the evaluation process. The culmination of the literature review process is a discussion on the results and how the key findings from the literature analysis guide both the experimental and the model development methodology, through a conceptual framework. Besides this, this review not only synthesizes existing work in this research domain but also sets the stage for a novel model that integrates advanced feature engineering, robust data preprocessing techniques, and state-of-the-art evaluation metrics to improve polarity shift detection. A summary, at the end, lays the groundwork for future research directions in the polarity shift detection. The contributions of this study aim to bridge the identified gaps in these key findings and advance the field of sentiment analysis in this research domain. The rest of the chapter is organized as follows: Section 2.2 describes the theory behind sentiment analysis. Section 2.3 discusses the architecture of sentiment analysis critically exploring all the processes involved. Section 2.4 discusses the factors affecting polarity shift in sentiment analysis. Section 2.5 reviews the existing sentiment analysis models for handling polarity shift, highlighting both their successes and limitations.

Section 2.6 discusses the key findings from the literature analysis and how they guided the empirical design. Section 2.7 discusses the conceptual framework, research variables and the operationalization of these variables. Lastly, section 2.8 discusses the summary of this chapter.

2.2 Theoretical Foundations of Sentiment Analysis

In the last decade, there has been an enormous growth in the availability of informal, opinionated texts like blog posts and product review websites. This has necessitated the growth of a field of sentiment analysis in order to address the question "What do people feel about a certain topic?" [3]. Khan et al notes that sentiment analysis has brought together various fields like computer science, computational linguistics, data mining, psychology, and even sociology[3]. This has witnessed the evolution of the traditional fact-based text analysis to enable opinion-oriented information systems. Sentiment analysis is underpinned by theories from various fields, particularly linguistics, psychology, and machine learning [28]. The subsequent sub section discusses the theoretical foundation of this research study.

2.2.1 Psychology Theory

Sentiment analysis extracts and analyses people's opinions, sentiments, attitudes, perceptions, toward different entities such as topics, products, and services. Psychology often taps into the ways human emotions and cognitive processes influence language use. Sentiment analysis employs a psychological theory called the appraisal theory[29]. The theory states that emotions arise from an appraisal of concrete situations that produce varying degrees of reactions on various people. These form a basis for structured sentiment extraction that is based on appraisal expression and a basic grammatical unit by which an opinion is expressed [29]. In psychology, the theory that positive language correlates with positive feelings and vice versa is critical for sentiment detection. It is evidenced that integrating psychological theory into sentiment analysis enhances both the interpretability and robustness of sentiment classification.

Psychology theory through appraisal suggests that emotions are the result of an individual's evaluation of a situation. In sentiment analysis, this theory supports context-sensitive models that consider discourse structure, user background, or goal alignment e.g sarcasm or disappointment. Moreover, psychological theories of language and cognition help explain how emotions are conveyed through text. It is clear that language reflects not just surface emotions but underlying cognitive appraisals such expectation violation. Example: “I expected better from this service” may not contain explicitly negative words, but signals dissatisfaction—a subtle sentiment inferred through appraisal theory. This study aims at detecting the implicit negation of mobile phone reviews therefore the psychology theory of language and cognitive underpins this research work. Psychology enhances the accuracy, nuance, and human-centric relevance of sentiment detection systems.

2.2.2 NLP Theory

Natural Language processing theory is the foundation displine that enables sentiment analysis by providing theoretical and computational tools that understand and manipulate human language [30]. Sentiment analysis, which focuses on extracting opinions, emotions, and attitudes from text, relies on NLP theories to interpret linguistic signals, model context, and resolve ambiguity for understanding and processing human language. In this process textual data is analysed to extract meaningful insights about emotions and opinions using various techniques and algorithms. Some of the NLP techniques and theories used in sentiment analysis are discussed in the subsequent sub sections.

2.2.2.1 Cognitive Linguistic Theory

Linguistics knowledge is very useful in sentiment analysis because it aids in the development of the corpus set. The corpus contains a large and structured collection of text data that is used for training, testing, or evaluating sentiment analysis models. Linguistic theory entails the ability to understand the characteristics of the data, determine which linguistic features may be applied and use ideas like “content” and “non-content” terms [31]. Linguistic theory entails the understanding of lexical semantic, syntax and pragmatics. Lexical semantic helps focuses on the meaning of words, including their positive, negative, or neutral polarity.

For example, words like "good," "bad," or "excellent" have positive, negative, or neutral connotations, respectively. Syntax helps in understanding the structure of sentences and how words are arranged which help to identify sentiment. For example, a sentence with a negative word in a positive context might be interpreted as sarcastic, highlighting the importance of syntax in accurate sentiment analysis. Pragmatics theory helps understand context in which language is used. These include; sarcasm and subtle nuances including implicit negation which can be tricky for traditional sentiment analysis algorithms. In relation to sentiment analysis, linguistic theory is very essential as it helps to explain how people convey feelings, attitudes, and opinions through language, and hence these can be detected and interpreted by computational models. Computational linguistics uses these models to obtain the computer-executable and understandable language from the vast source of raw languages textual data through proper representation and extracts the associated sentiments [32]. Therefore, it is evident that computational linguistics forms the foundation for performing sentiment analysis.

2.2.2.2 Text Representation Theory

Text representation is crucial in sentiment analysis as it transforms textual data into a format that machine learning models can understand. Machine learning algorithms require numerical data as input, and text representation converts text into this format. In sentiment analysis effective text representations should capture the meaning and relationships between words, thus enabling models to better understand the sentiment expressed in a text [33]. The text representation theory allows sentiment analysis models to achieve higher accuracy in identifying the emotional tone of a text. This theory incorporates a number of techniques to perform this task [34]. These include; Bag-of-Words, TF-IDF, and word embeddings. Bag of words technique considers only the frequency of in a document but ignores word order. TF-IDF technique assigns weights to words based on their frequency in a document and how rare they are in the entire corpus. Word Embeddings are for dense vector representations of words that capture semantic relationships between them. These include Word2Vec, GloVe and contextualized embeddings like BERT that considers the context of words within a sentence or document to create more nuanced representations.

2.2.2.3 NLP Techniques

NLP theory support a number of NLP techniques which include; data preprocessing techniques and feature extraction techniques in sentiment analysis [35]. Data preprocessing techniques include tokenization, stop word removal, stemming and lemmatization. Tokenization breaks down text into smaller units (tokens) like words or characters, which is a foundational step in NLP and sentiment analysis. Stop word removal eliminates common words (like "the," "a," "is") that don't carry much semantic weight, making the analysis more focused. Stemming and lemmitazation involves reducing words to their root form (e.g., "running" to "run"), improving the accuracy of sentiment analysis [36]. Feature extraction techniques include Bag-of-Words (BoW) and TF-IDF. Bag-of-Words (BoW) and TF-IDF are the common feature extraction techniques that convert text into numerical representations suitable for machine learning models. Word Embeddings (e.g., Word2Vec, GloVe) represents words as vectors, capturing their semantic relationships and allowing for more nuanced sentiment analysis.

2.2.3 Machine Learning Theory

The theoretical principle of machine learning is motivated by the need to improve accuracy, contextual understanding, and generalizability across sentiment analysis tasks. Sentiment analysis is defined as a procedure to identify, recognize, and/or categorize the users' emotions or opinions for any service as positive, negative, or neutral [29]. According to research, even though sentiment is stated as a polarity in computational linguistics, it is typically treated as a classification task [5]. Sentiment analysis model learns patterns from labelled training data that consists of text labelled with sentiment categories (positive, negative, or neutral). Classification algorithms play a central role in sentiment analysis by learning to predict sentiment polarity from examples of text. A number of machine learning theories that underpins sentiment analysis are discussed in the subsequent sub sections.

2.2.3.1 Supervised Learning Theory

Most sentiment analysis models use labelled examples such as positive, negative, neutral to guide model learning therefore are termed as supervised classification tasks [37]. Based on the statistical learning theory, supervised classification tasks aims to approximate text embedding to sentiment labels. The key concepts in

supervised learning are empirical risk minimization where models are trained to minimize loss on training set. Generalization bound provides a gap between performance on training data and unseen data using VC dimension and Probably Approximately Correct learning, which relate to model complexity and data size.

2.2.3.2 Representation learning

In machine learning, representation learning states that effective models can automatically learn meaningful abstractions or features from raw data [38]. In sentiment analysis, traditional machine learning feature engineering is very key to enhance the model's classification. Whereas in deep learning models like the transformers learn dense distributed representations that encode syntactic and semantic relationships relevant to sentiment.

2.2.3.3 Optimization Theory

The optimization theory supports the training process of the sentiment classification models inorder to minimize the loss function. Gradient descent algorithms such as SGD and Adam are used in deep architectures for loss minimization [39]. On the other hand, Convex and non-convex optimization are used with traditional models. It is therefore critical to understand convergence, learning rates, and load points that are key in the training of sentiment classifiers.

2.2.3.4 Regularization Theory

This theory addresses the challenging problem of overfitting, which usually occurs in deep learning architectures. It prioritizes the learning of machine learning rather than memorizing through reduction of the generalization error [40]. Most deep learning models for sentiment analysis often face the risk of overfitting. In this study regularization theory helped mitigate this by penalizing model complexity using drop out and early stopping.

2.2.3.5 Hybrid Learning Theory

Hybrid learning theory asserts that combining multiple learners improves prediction accuracy. This is well explained by the no free lunch theorem in machine learning which justifies that no single algorithm can handle all the problems [39]. In sentiment analysis, this is implemented through the hybrid architectures through combination of two or more algorithms inorder to capture diverse linguistics.

In this study, the sentiment analysis model is implemented through combination of transformer, LSTM and attention mechanism.

2.2.4 Information Retrieval and Text Mining Theory

Sentiment analysis is a subfield of Natural Language Processing (NLP) that heavily draws upon the theoretical foundations of Information Retrieval (IR) and Text Mining. The information retrieval and text mining provide essential techniques and models for preprocessing, representing, organizing, and extracting meaningful insights from large volumes of textual data [41]. They are essential in sentiment analysis in order to enhance the ability to identify, classify, and quantify opinions expressed in unstructured content such as reviews, social media, or survey responses. Information retrieval theory reinforces how text data is transformed into structured representations. In sentiment analysis, the Vector Space Model (VSM) like TF-IDF and word embeddings represents documents or sentences as vectors where each dimension corresponds to a term or feature. This representation is very key for computing similarity, distance, and relevance, which are very relevant in sentiment classification. In this study, text preprocessig, feature engineering, text classification and pattern discovery was underpinned on this theory of information retrieval and text mining.

2.3 Architecture of Sentiment Analysis Models

Based on the scope of sentiment analysis, the sentiment analysis models can be implemented at various granular levels. At present, there are mainly four levels of sentiment analysis that have gained popularity. They are document level, word level, Sentence/Phrase level and aspect level[5]. The document level sentiment analysis is performed on a whole document and a single polarity given [21]. In Sentence level analysis, each sentence is analysed and a corresponding polarity is found. Researchers have recommended this when a document has a wide range and mix of sentiments associated with it. The aspect level sentiment analysis is performed at each aspect in the sentence. In this the classification primary attention is to all the aspects used in the sentence and a polarity is assigned to all the aspects. Later an aggregate sentiment is calculated for the whole sentence in classification level.

The word level investigates the polarity of each and every word and investigates how impactful individual words have on the overall sentiment.

According to research, sentiment analysis models have progressed from simple rules based models to advanced deep learning techniques [5]. These sentiment analysis models have adopted a general framework of sentiment analysis as shown in figure 2.4. The framework comprises mainly four steps; collection and standardization of data; pre-processing of the dataset; extraction of features; classification of the sentiments and an optional step of sentiment standardization.

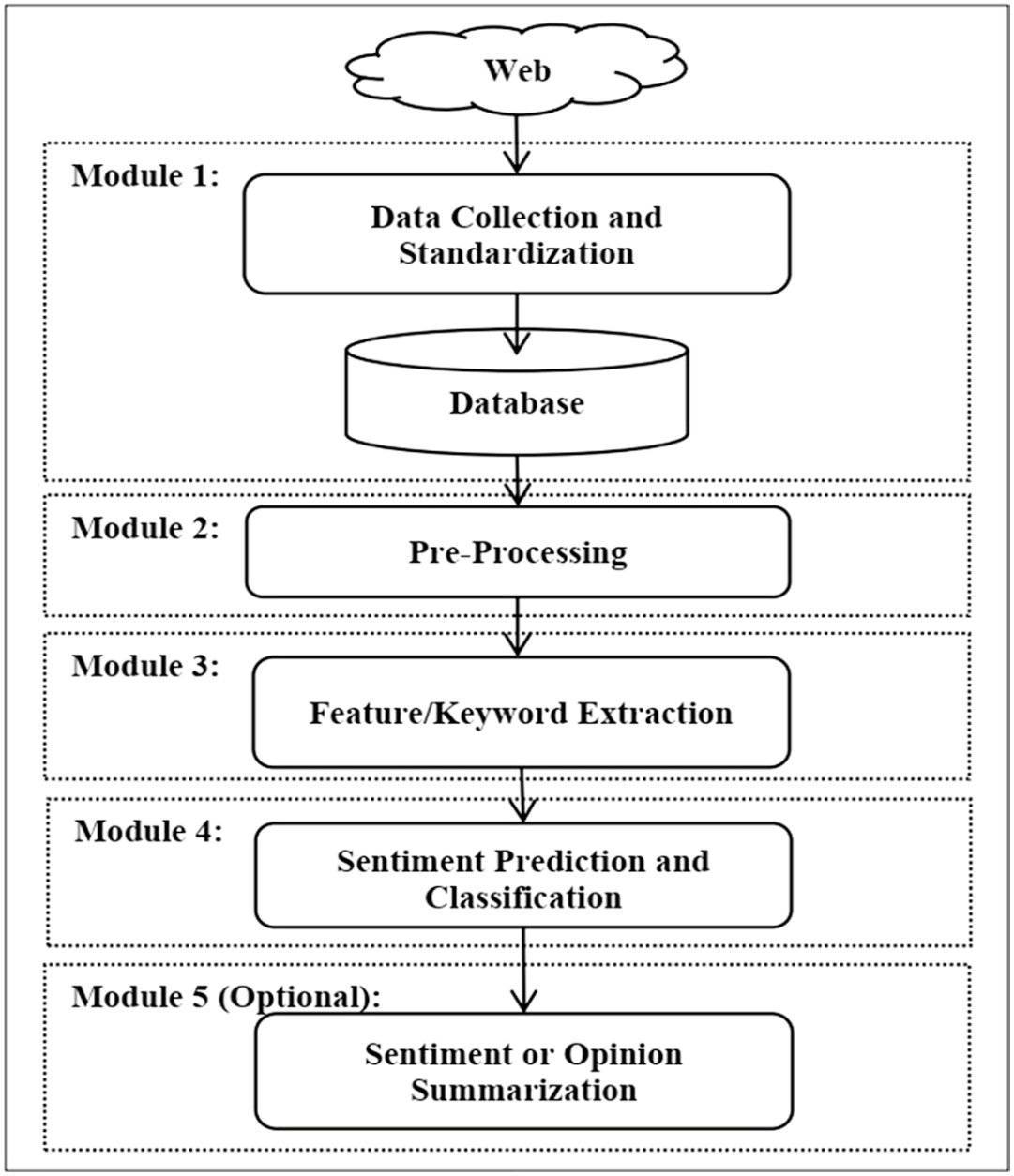


Figure 2.1: Sentiment Analysis Framework [5]**:**[42]

2.3.1 Data Collection and Standardization

Due to the availability of various platforms of expression, the type and format of people's views, opinions, or sentiments on a particular subject is increasing. Among the different available types of data such as text, image, audio, or video, the research on textual data has been appreciated in the last few years**.** It is also evident that though multi-lingual text data has attracted few researchers, however, 90% of sentiment analysis studies concentrate mainly on English textual data [5]**.** The development and validation of a sentiment analysis model typically depends on the quality and the structure of the dataset from the boundless and voluminous to a great extent. Bordoloi and Biswas [5] identifies the following sentiment analysis datasets that are publicly available; Blitzer's multi-domain sentiment data, Public product reviews, UCI Machine Learning Repository, and Sentiment Labelled Sentences. The publicly available mobile phone reviews dataset include; Amazon reviews devices like mobile phones, laptops, televisions, tablets, and video security systems used by Patra. Their proposed BERT model was used for sentiment categorisation and it demonstrated ability to capture long-term dependencies in textual data [43]. Sivakumar et al [44]used various datasets to classify customer review sentences using an intelligent system using long-term short memory with fuzzy logic. These included; Amazon cell phone review, Amazon video games review and consumer reviews of amazon products benchmark dataset [44]. Alharbi et al. have used 400 K reviews of unlocked mobile phones sold on Amazon.com which is publicly available on Kaggle website [45]. Lastly, Imtiaz & Ben, highlights the five popular websites where mobile phone datasets is publically available [46] .These include: www.amazon.in suggested as one of the largest e-commerce website around the world with a rich mobile section with a huge number of consumer comments.

2.3.2 Data Pre-processing in Sentiment Analysis

This is the process of preparing a clean, relevant and well-structured dataset for sentiment analysis through removal of any sort of noise. In the preprocessing stage, the raw text data undergo cleaning to remove irrelevant information such as stop words, punctuations, special characters, short words and numbers [47]. This process aids in text extraction because some symbols, URLs, and words are not necessary for classification features.

At the end of this process normalized text data is produced that is used for sentiment classification. The data cleaning steps are shown in figure 2.2.

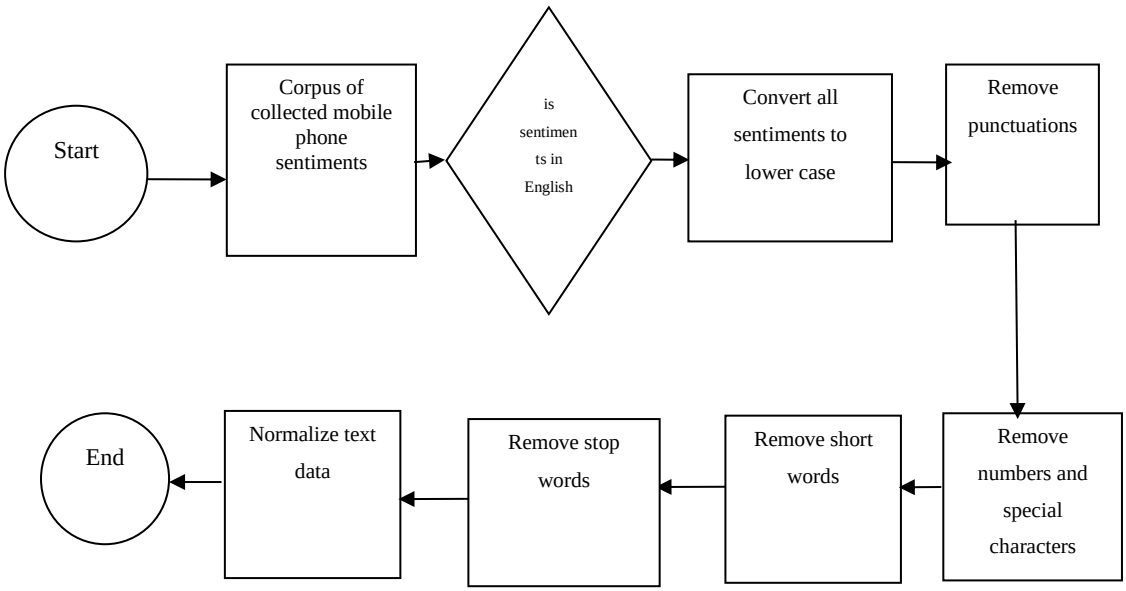


Figure 2.2: Flow Chart of the Data Cleaning and Preprocessing [48]

The data cleaning process begins by the analysis of corpus containing mobile phone sentiments. Sentiments are normalized by ensuring all text is written in English and changing them to lower case. Removal of all irrelevant information including punctuations, numbers, special characters, short words and stop words is done. Finally the output is normalized data that can be used by the machine learning models.

Researchers have noted that preprocessed data has an effect on the final sentiment classification accuracy. Researchers [47]; [49]; [50] have investigated the effect of pre-processing task by employing various text preprocessing techniques that include: removal of stop words, removal of short words, word stemming, word lemmatization and tokenization on the datasets[51].They concluded that using a good text preprocessing technique is critical to the sentiment classifier’s performance for unstructured reviews. Some of the popular preprocessing techniques in sentiment analysis are discussed in the sections below.

The removal of stop words is a data preprocessing that seeks to get rid of "functional words "words like "this," "is," and "but" that don't have any relevance in the text. Because they frequently appear in texts and have little impact on classification,

the stop-words have little effect [52]. The SMART stop-list, a set of frequently used English stop words, is employed in this investigation. Before reading each stop-word individually from the SMART stop-word list, the tokens are first organized into an array. The token is then matched against the stop-word and eliminated if they match.

Removal of short words is a data preprocessing that helps to reduce a number of tokens or characteristics, a brief word of one to two letters in length is removed [53]. Short-form words frequently transform into such short words as "tq" (thank you) and "ok" (read: okay). The grammar and spelling errors are as well considered as noise since they frequently appear in informal texts [53]. To reduce the amount of features in the text structure, elimination is required.

Word stemming data preprocessing is the process where the root word is sorted by eliminating suffixes, which decreases the token. Different tokens with the same root-word in a feature space are identified as the same token. In this regard, the words "continued" and "continuing" have the same root as "continue". Furthermore, stemming is done for both singular and plural words. In this case, the word "levels" comes to "level". In this research, Porter's method for English sentences, based on a five-step process, is applied. Porter's algorithm is helpful and quite accurate, thus it is commonly utilized for classifying texts for a number of applications [54].

Word lemmatization preprocessing involves conversion of word by using dictionary lookup and linguistic rules so as to remove inflectional conclusions and replace them with a word's base or dictionary form [54]. In machine learning lemmatization tools like WordNetLemmatizer and spaCy available in the python's natural language toolkit are used. This process improves the data through standardization of words, minimization of noise, and dimensionality reductions. Consequently, this improves the accuracy of sentiment classification models. The work of Alam and Yao [55] on the effects of preprocessing of data on the sentiment classification accuracy shows a significant difference. They conducted comparison with and without lemmatization process and the results showed a significant difference.

The works by Haddi et al. [56] shows improved sentiment classification accuracy when lemmatization data preprocessing technique was carried out. Figure 2.2 shows an example of lemmatization process of mobile phone reviews.

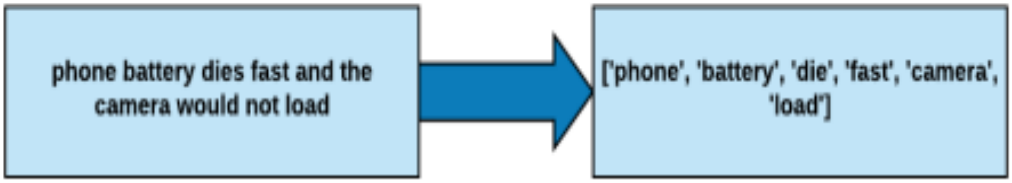


Figure 2.3: Lemmatization of the Mobile Phone Reviews [57]

The process of tokenization involves breaking down chunks of text into meaningful words, or tokens [58]. A text document's performance is also not affected by hypertext mark-up language tags, XML scripts, special characters, or punctuation. Therefore, getting rid of them helps to greatly reduce down on the number of features needed for classification. Tokenization process is done after conversion of text to lower case, removal of punctuation and elimination of characters are done.

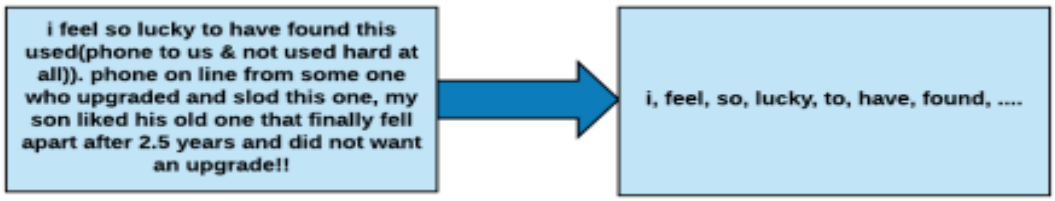


Figure 2.4: Tokenization Process of the Mobile Phone Reviews [47]

Despite the cleaning of the dataset using various preprocessing steps, using all of the data in the dataset can result in dimensionality issues, longer computation times, and the use of irrelevant or less significant features or terms. The subsequent section discusses the feature selection techniques used in sentiment analysis and their impacts in the classification.

2.3.2.1 Feature Selection

Feature selection is a very useful preprocessing step in sentiment analysis, it aims at choosing a suitable set of features in order to improve the models' effectiveness [58]. A set of features from meaningful words obtained from data are used in sentiment analysis. The selection of subsets is a process for narrowing down a large corpus's relevant characteristics with the goal of enhancing classification efficiency [59]. With feature selection techniques, sentiment analysis algorithms execute faster, making predictions and data analysis process more accurate. Feature selection also helps in minimizing the dimensionality of the feature space through the identification

of the most important features [60]; [61]. Feature selection in sentiment analysis has been classified as filter, wrapper, or hybrid approaches.

Filter based approaches use statistical measures to rank the features where the highest ranking is selected. Firstly, each feature is assigned a score while the top-scored features are then selected and ranked [62]. According to Bommert et al.,[50] the filter based approach is suitable for datasets with a high number of features besides working independently from the classifier used. Some commonly used filter based approaches are Information Gain (IG) [63], Chi-square (CHI) [61], Document Frequency (DF)[61],and Mutual information [61]. A number of researchers have adopted the filter based approach to improve sentiment classification. Nurhayati et al. conducted a study to find out how the Naïve Bayes algorithm performed when Chi-square feature selection was used. According to the results, Chi square improved the recall, accuracy, and precision of the Naïve Bayes sentiment classification. In the studies conducted by Paudel et al.,[64] and Falasari and Muslim, [65] the researchers concluded that Chi-square could increase the efficiency of the computation process while also increasing the accuracy of the classification method. Manek et al., [66] in their research used a feature selection technique based on Gini Index that was evaluated with Support Vector Machine classifier. The researchers [67] used particle swarm optimization in their study in order to improve the classification outcomes of the NBC algorithm, also known as the NBC-PSO algorithm, which operated on the principle of increasing feature weighting. Elangovan and Subedha [68] suggested the Firefly (FF) and Levy flights (FFL) models in their study as an approach to extract features from a web-based review. The model selected features which are optimal within a high dimension feature subset. Researchers have also adopted wrapper based feature selection approaches. This technique combines machine learning algorithms with optimization of search algorithms in order to identify the most crucial features [61]. The wrapper based approach evaluates the features chosen based on the classification accuracy using a function called fitness. According to Tubishat et al,[63] the wrapper-based methods improve classification besides reducing the original feature space. A number of researchers, in their works, have applied the wrapper-based methods for the feature selection. The researchers [67] adopted Particle Swarm Optimization (PSO) in their study in order to enhance the classification outcomes using the Naïve bayes

classifier. Kristiyanti and Wahyudi compared three wrapper feature selection to enhance the classification performance of the Support Vector Machine i.e. the Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and the Principal Component Analysis (PCA). In their comparative study the PSO performed better than GA[69].

Hybrid feature selection approach combines the use of filter and wrapper based methods in order to obtain the best possible feature subset [54]. Syafiqah and Suryanti suggest an enhanced machine learning-based hybrid feature selection method that improves the sentiment classification accuracy[54].They employed recursive feature elimination, word frequency-inverse document frequency, and support vector machines for sentiment classification [54].The SVM score and recursive feature elimination were utilized to evaluate and rank the features through the identification of the most important features. Elangovan and Subedha, suggests the Firefly (FF) and Levy flights (FFL) models in their study as an approach that extracted features from reviews obtained from a website [68]. Patcharanikarn and Wararat, suggests use of a hybrid of the Information Gain and Synthetic Minority Oversampling Techniques (SMOTE) for feature selection in order to obtain appropriate features for each class [60]. It is very evident that the hybrid feature selection model increases the classification accuracy. However, it is pointed that the characteristics of the dataset and the exceptional difficulties in managing polarity shift in the sentiment analysis task should inform the choice of feature selection. It is therefore critical to experiment with multiple feature selection techniques and evaluate the effectiveness in improving model performance in handling polarity shift in the sentiment analysis.

2.3.2.2 Data Augmentation Preprocessing

According to research, the original data augmentation is used in image classification by increasing image data such as rotate, translate, scale, add noise, etc., [17]. Recently data augmentation has been applied in sentiment analysis to automatically generate more training data using various techniques [85]. It entails the generation of new data from the original data so as to enrich training data without user intervention [17]. In text, data augmentation is more complex and therefore many studies have investigated the techniques to get new data, as well as improve the quality of new

data without user intervention. This process enhances the original training data and hence increases accuracies of sentiment analysis models. Back translation is one of the data augmentation techniques that aim at obtaining more training samples though use of the translators to translate the original data to a certain language. Then the translated data is passed into the independent translator to translate back to the original language. It is notable that English is one of languages having many training datasets for translation because others lack training datasets for translation models [85]. Sennrich et al. [12] and Sugiyama and Yoshinaga [13] used back translation technique to generate training data in order to improve performance of classification model. Sugiyama and Yoshinaga, [13] used back translation technique to improve performance of classification model by generating more training data. Hou et al. [86] used sequence-to-sequence model to augment the training data by generation of lexical and syntactic alternatives.

Synonym replacement obtains random synonym words and replaces them for a new sentence. Kobayashi [87] proposed the contextual augmentation to stochastically replace words with other words. Random Swap allows for a swap n times two non-stopword words randomly. Random Insert finds a random synonym of a non-stopword word and insert this randomly n times in the sentence. Random Delete randomly removes each word in the sentence with probability p. These processes have repeated many times until having the expected training data. The biggest disadvantage of these methods is that they do not reserve the contextual meaning of the sentences.

2.3.3 Feature Extraction Process in Sentiment Analysis

In a sentiment analysis model, the words and symbols within the corpus are mainly used as the features. The object feature has been used in sentiment analysis to indicate the feature of an object. Let d be a document, containing reviews about the products. The opinionated document is a collection of expressed opinion by user in statements. Hence, document d can be described with the help of a series of sentences **d = <s1, s2, . . . , sm>.** Feature extraction is the process of transforming random text and images into numerical that could be used by sentiment analysis models [53]. Since the mined user reviews are mostly in the form of text, there is need to represent them numerically in order to be processed by the sentiment

analysis algorithms. During the feature extraction process in sentiment analysis, a document is treated as a Bag of Words (BOW) that is projected as a feature vector, and then later categorized using a classification technique. During feature extraction, the keywords are mainly represented using either a Vector Space Model (VSM) or a Graph-Based Model (GBM).

Vector space model (VSM) represents the documents as vectors of the terms [5]. The model builds a matrix V which is termed as a document-term matrix where the rows represent the document in the dataset, whereas columns correspond to the terms of the whole dataset. Thus, if the set of documents is represented by D = (d1, d2, ..., dm) and the set of terms/tokens representing the entire corpus is T = (t1, t2, ..., tn) , then the element dti,j ∈ Vmxn , i = 1, 2,…,m , and j = 1, 2,…, n is assigned a weight wi,j . The weights can be assigned based on the word frequency associated with a document or the entire dataset. Word embeddings clusters words with similar meaning together in a vector space. Each word is assigned to a vector, which is then learned in a manner similar to neural networks. Word2vec creates word embedding through the use of a 2-layer neural network for vectorizing the tokens [5].

Graph based model uses a node or vertex Vi to represent a document term or function ti and the edges Ei,j represents the relationship between them [5]. Nasar et al., [70]showed that various properties of a graph, like centrality measures, node's co-occurrence, and others, play a significant role in keyword ranking. Semantic, syntactic, co-occurrence, and similarity relationships are some of the specific perspectives of graph-based text analysis. The importance of a term is calculated by using the centrality measure, to calculate the importance of the node in the graph [9]. A number of vectorization techniques, discussed in the subsequent section, have been adopted by various researchers in sentiment analysis.

2.3.3.1 TF-IDF (Term Frequency and Inverse Document Frequency) Approach

The term frequency and inverse document frequency score is used to highlight key phrases in a written form. It weighs the importance of a word based on their occurrence in a document as well how rare the word is within the corpus matrix [71]. This approach tries to reduce the influence of irrelevant words in a document [54]. The term's phrase frequency (tf) is clearly defined as the frequency of a term in

relation to the overall word count of the text. The IDF (Inverse Document Frequency) formula is used to determine the significance of a phrase [54].
The TF-IDF score is computed using the following equation.

$$tf - idf = tf * log\ log\ N\ \ DF + 1! \ (1)$$

where,

The term "tf" stands for the phrase count in the article.
The initials df stands for the number of data points which include a word.
There are news articles overall, totalling N.
Inverse Document Frequency (IDF) formula is formulated as follows:

$$IDF\ (t) = log\ log\ (\frac{N}{DF})$$

The number of documents including the phrase "t" is represented by DF, while the total number of documents is represented by N.
Using TF-IDF, the textual representation of data can be converted into a Vector Space Model (VSM).

Japhne and Murugeswari, in their study on opinion mining complex model for handling polarity shift, uses the TF-IDF technique to determine the important word [72]. To achieve this, the technique considers the number of documents that the word appears. Ahmad and Aftab, in their work, uses TF-IDF to evaluate the frequency of the most useful words in the document[73]. The process makes their sentiment detection process easy. Kumar & Garg, in their empirical study, uses TF-IDF to extract the most important terms in a text [15]. They also assessed the importance of a word in the corpus using the TF-IDF vectorizers. These are later used as the input features for training the classifier. However, Alzami et al. compared the effects of the various feature extraction methods in relation to the performance of the sentiment classification models. They compared Word2Vector, Word Bags, TF-IDF, and a hybrid of TF-IDF and Word2Vector [74]. It is noted that the main drawback of TF-IDF is its inability to capture text position, text semantics and co-occurrence in many documents.

2.3.3.2 Use of Bag-of-Words

The bag-of-words(BoW) method in NLP represents a text document as an array of fixed-length vectors [75]. It represents each document as a vector of word frequencies, typically using a fixed vocabulary [76]. The vector's length is equal to the document's vocabulary size. This fixed length vector in the input document holds a value dependent on the word's occurrence [75].The technique employs fixed-length vector format, making it appropriate for classification and clustering problems involving classification. BoW technique proves to be an ineffective model for classification since it doesn't care about the semantics of the sentence. The use of BoW as a feature extraction method has been applied in a number of sentiment analysis models used in handling polarity shift problems. Thevar et. al. created a sentiment analysis model for handling polarity shift using the BoW feature extraction method [77].

CBoW uses key words in the context of predicting a word. Appropriate words are used as the input, and they are utilized for predicting the desired term. The corpus's vocabulary must be created in order to create a CBoW model [23]. Target and context generators are created afterwards, and inputs are passed into the embedding layer before the lambda and dense layers. It is noted that continuous bag of words guarantees all the vectors for each individual term orthogonally and shares all contextual information. It is therefore most appropriate for neural network-based approaches [67]. On the other hand, it provides contextual information in addition to representing a word as a vector.

The continuous bags of words methods are applied to a number of sentiment analysis models used in handling polarity shift problem. Liu, in his study, uses continuous bag of words (CBoW) for vector representation of text [77]. For text sentiment analysis, feedforward neural networks serve as the foundation for the CBoW language model.

2.3.3.3 Word Embedding

This feature extraction technique represents words as dense vectors [78]. Some examples of word embedding include; Word2Vec, GloVe, and FastText. Word2Ve uses a shallow two-layered neural network of the word embedding model [79]. A

word2Vec embedding is the term used to describe a word's numerical representation as a vector. Word2ec takes into account the related semantics of words to produce embeddings including the context of the word in a text. In the same or similar circumstances, the related words are mathematically classified into a vector space [79]. Avinash and Sivasankar, in their study, used Doc2vec which is an extension of word2vec and doc2vec model [80]. Doc2vec improves the word-to-word sequences learning process of the embeddings that represents documents as fixed-length vectors with low dimension. Singh and Paul, in their study, used word and character embedding for feature extraction. They conducted pre-training of the word embeddings in order to get the semantic and syntax information for each word in a sentences [81].

Additionally, character embedding is used in order to capture negations that are morphological in nature. As this usually assesses the frequency of helpful phrases, this offers significant and valuable information during the pre-processing stage and ultimately simplifies the sentiment detection process. The identification of significant information is highly influenced by the term frequency, as earlier explained by [82]. Deho et al. generates high-dimensional word vectors that can recognize word context using Word2Vec [83]. The results of the word vectors are then used to train the models. Their results showed improved classification accuracies. Rajabi et al. uses word embedding model to represent text semantically [84]. They uses batch number as per the concepts from the “ConceptNet” semantic network. Arora and Kansal proposes a deep convolutional character level embedding for text normalization.

Sentiment analysis is performed using a neural network model [85]. This model performs well when managing the small memory space difficulties in word level embedded learning as well as in the processing of noisy sentences. The unstructured data performs well with this approach. Xia et al. suggests a character level embedding as a solution to the issue of the current text sentiment analysis algorithms' inability to extract sentiment words. Additionally, it resolves the pre-training word vector issue caused by the vocabulary gaps [86].

Singh and Paul identifies lack of contextual information as the main problem associated with one hot encoding vector [81]. For instance the phrases "have a nice day" and "have a great day". Regardless of the fact that the two sentences are fairly similar, they are unable to find any similarities between them when utilizing one hot encoding approach. This is because of the fact that each term's vectors are orthogonal and do not share any contextual information. Rezaeinia et al. in their study indicates, that among the most efficient and practical word embedding feature extraction techniques are the Word2Vec and GloVe. Nevertheless, they point out that ,these approaches overlook text sentiment information, and that in order to produce precise vectors during training, a sizable corpus of texts is required [87].

2.3.3.4 Count Vectorizer

This involves simple vectorizers that convert every data token into a feature, through simple vectorization. Each text in the document is taken into consideration, and each relevant ID and word repetition is taken into account [75]. The classifier is able to accept this multiset of words as input. The dataset is referred to as sparse because each file has a large number of zeros for every text that is absent from the dataset [75]. The count vectorizer methods have been applied in a number of sentiment analysis models used in handling polarity shift problems. Rajat et al., in their study, carried out a sentiment analysis on a sizable real dataset from Amazon using the term frequency inverse document frequency and the counter vectorizer [88]. Irawaty et al., in their study, conducts a comparison of the effectiveness of the following vectorizers, namely TFIDF Vectorizer, Hashing Vectorizer, and CountVectorizer [89]. Their results show that the best vectorizer is TFIDF because of its relatively higher accuracy in the prediction of the negative values. Additionally, there are more positive predictive values as compared to other vectorizers. However, Kalaivani et al., in their review, notes that count based embeddings does not exactly represent syntactic and semantic words in a document [90].

2.3.3.5 N-grams

N-grams, also referred to as bi-grams or tri-grams are used to capture word combinations that carry sentiment [91]. The N-grams feature extraction methods are applied to a number of sentiment analysis models used in handling polarity shift problems. Ahuja et al. in their study, compares the N-Gram feature extraction

method and the TF-IDF word level on the SS-Tweet dataset [91]. For classification, they employed the Naïve Bayes machine learning algorithm, decision trees, support vector machines, and logistic regression. They came to the conclusion that using TF-IDF word level improves sentiment analysis performance by 3-4% when compared to N-grams. Kaur et al., in their study, used N-gram to perform feature extraction and post tagging on the sentences [92]. They used the kNN classifier for sentiment analysis. Mutinda et al., in their work on sentence level sentiment analysis uses a fixed hybrid N-gram window for feature extraction and a minimum redundancy maximum relevance feature selection algorithm [93]. Their methodology enhances the N-gram feature extraction method by creating a hybrid vector from words, semantic orientation of the words, and Part of Speech (POS) tags.

The most widely used feature extraction methods in sentiment analysis are bag of words, N-gram, and Term Frequency-Inverse Document Frequency [93]. However, it is noted that these methods face difficulties when they fail to take word relationships into account. They have the high feature dimensionality issue in addition to disregarding the properties of words. Additionally, it is notable that Word2Vec, Glove, FastText and BoW are the most prevalent feature extraction methods [94]. It is noted that the main disadvantage of Word2Vec and Glove is their ability to generate a random vector in a word, not in the dataset. FastText on the other hand is a continuation of Word2Vec that can overcome this disadvantage.

2.3.4 Sentiment Classification Process

Researchers have suggested different techniques for sentiment classification purposes [47]. Duong and Nguyen presented different classification techniques of sentiment into machine learning and lexicon based approaches. The machine learning approach of sentiment classification uses well-known machine learning to classify linguistic features into appropriate sentiment classes [95]. Sentiment classification techniques are classification into either machine learning approach that entails supervised, semi supervised, unsupervised and self-supervised learning or lexicon-based approach that entails either dictionary or corpus based approaches as shown in figure 2.5.

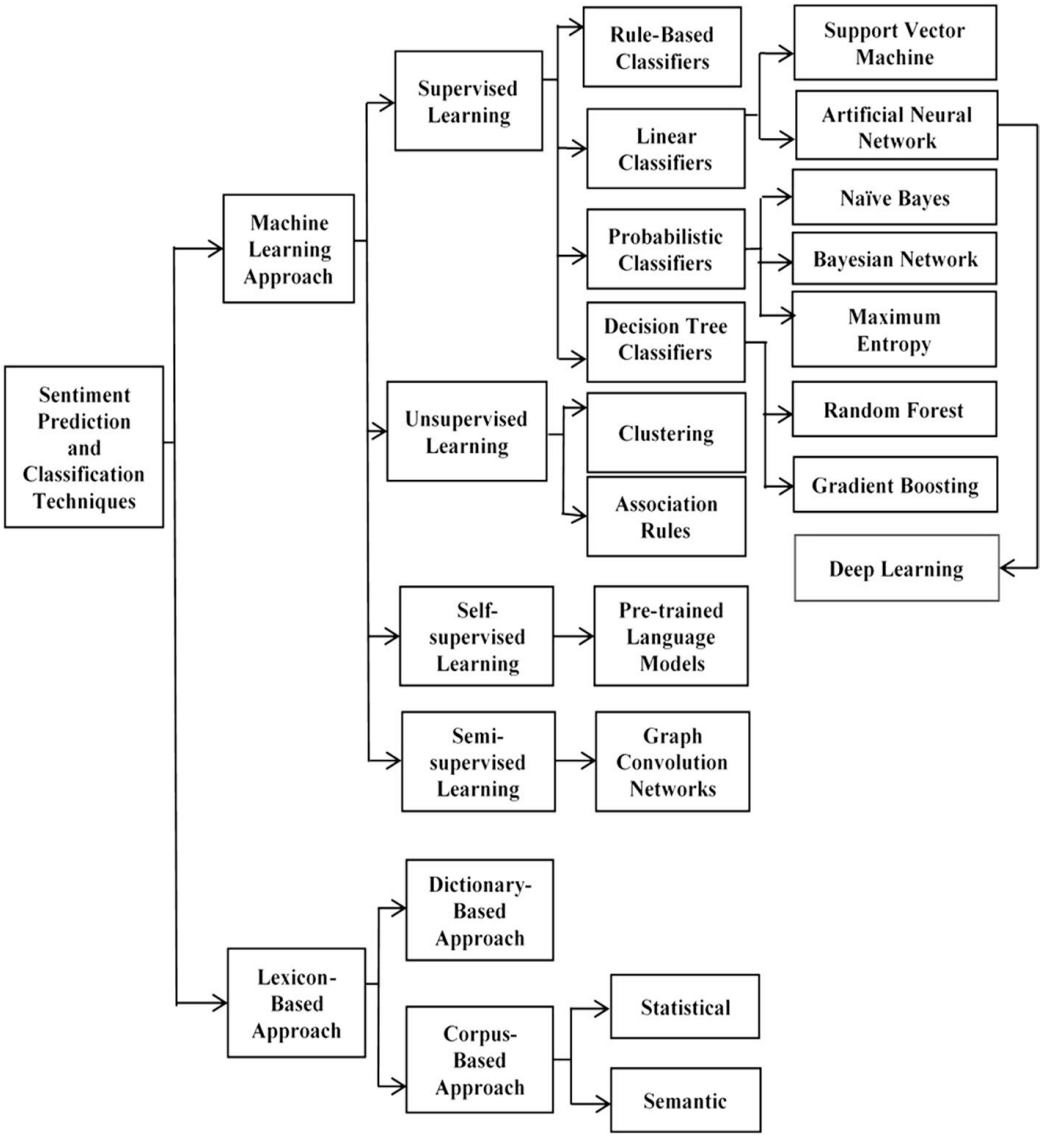


Figure 2.5: Sentiment Classification Techniques [5]

The supervised machine learning approach uses the initial set of labelled opinions to determine the associated sentiment or opinion of any document. Deep learning uses neural network layers capability to handle large datasets and their associated complexities such as non-linearity, intricate patterns.

The different supervised learning techniques employed in the sentiment analysis process include Naive Bayes, Support Vector Machines (SVM), Decision Trees, Random Forests, and Maximum Entropy, [96]. Deep learning structure allows for the transformation and extraction of features automatically, which facilitates self-learning as it goes by multiple hidden layers, in a way similar to humans.

At present, deep learning such as long short-term memory (LSTM), gated recurrent unit (GRU), bidirectional long short-term memory (BiLSTM) or convolutional neural network (CNN) are prevalently preferred in sentiment classification [22]. Moreover, Yadav and Vishwakarma [96], in their study on unsupervised learning, the models are trained using unlabeled datasets. Most of these techniques rely on clustering methods such as k-means clustering, expectation-maximization, and cobweb. Self-supervised learning trains the model with unlabeled datasets and then trains itself to learn a part of the input by leveraging the underlying structure of the data. Pre-trained Language Model (PML) is one of the self-supervised learning techniques which have been adopted in sentiment analysis [97]. The main challenge with self-supervised learning is that they pay less importance to the sentiment-specific knowledge during the pre-training phase.

Lexicon-based approach uses opinion terms to execute sentiment classification. This method uses the notion that a sentence's or document's cumulative polarity is the sum of the polarities of individual terms or phrases [21]. Dictionary-based approach has been adopted in lexicon-based approach, it starts with a series of manually annotated opinion terms. The collection is then extended by searching through a dictionary to find synonyms and antonyms. Corpus-based approach based has also been adopted in lexicon-based approach, it uses syntactic patterns or co-occurring patterns in a vast corpus to extend the original seed list of opinion terms [95].

2.3.5 Sentiment Summarization

Sentiment summarization also known as opinion summarization is the process that provides an idea of the overall influence or polarity depicted by the dataset. The process sums up the polarity of all individual words/aspects /sentences/documents of the dataset. Sentiment summarization is meant to provide a generalized polarity depicted by the whole dataset [5]. The standard set for sentiment summarization uses the following assumption; a set of documents D = {d1, . . . , dm} that contain opinions about some entity of interest. The main aim is to generate summary S of that entity that is representative of the average opinion. Li et al. [98] notes that the existing sentiment-based text summarization methods only considers the sentiment polarity of either a single sentence or a whole document.

This approach ignores the changes of sentiments along with sentences or sentiment flow across the whole document. The authors propose a topic sentiment summarization framework based on reaching definition (TSSRD). Firstly, they use topic models to model documents and calculate topic sentiment embeddings. Secondly, the document structures analysis is done from different perspectives. The improved reaching definition is used to analyse sentiment changes and sentiment flow. Finally, topic sentiment summaries are generated based on sentiments in steady states of the reaching definition [98]. Researchers have concluded that both review summarization and sentiment classification can be combined to improve sentiment classification accuracy [99];[98].

2.4 Factors Affecting Polarity Shifts in Sentiment Classification

According to research polarity shift is identified as one of the major problems with the existing sentiment analysis models. The main causes of polarity shifts are referred to as polarity shifters. Polarity shifters are the factors that can change a word's prior polarity in one of three ways i.e. rise, decrease, or neutral [23]. According to Madhuli and Rahuli, there is need to consider the most significant factors that are key to polarity shifts [100]. The effectiveness of the sentiment classification models can be enhanced by the automated detection and handling of such factors. If not detected and handled effectively, these polarity shifters can have detrimental effect on the classification performance of these models [23]. The existing polarity shifters include negation, contrast, context, ambiguity, evolving language, subjectivity, cultural and regional differences, news and events, authors tones, review fraud, pragmatics and domain-specific knowledge [101]; [26];[102]. The presence of negation words like "not" and "never" can invert the sentiment. For example, "not bad" means "good" [81].

Negations words like “not”, “neither”, and other similar phrases can change the polarity of a sentence, thus making them crucial for sentiment analysis [23]. Positive expressions include "this movie is good" and negative ones include "the movie is not good". In most times, certain approaches eliminate negation terms since they are termed as stop-word lists or disregarded in entirety [81]. The polarity shift of a sentence is based on two essential components of negation.

These include the cue and the scope. The covering of negation's effects on the sentence is its scope [81]. According to Nikhil, classification of negation cues is done into three different categories i.e. linguistic negations, structural and valence shifters [103]. Negation in linguistics can either be explicit or implicit. While implicit negation uses phrases like rarely, avoid, etc., the explicit negation on the other hand expresses itself clearly [76]. Sarcasm and irony are forms of figurative speech where the speaker means the opposite of what they say. Detecting sarcasm and irony in sentiment analysis can be challenging for sentiment analysis models, resulting in polarity shifts problems [104]; [105]. The sentiment expressed in a statement can change depending on the context. For example, a phrase that might seem positive in one context can turn negative in a different context [105]; [102]. Ambiguous language or statements, interpreted in multiple ways, may lead to polarity shifts, as different interpretations can carry different sentiments [84].

Language evolves over time, and the meaning of words or phrases can change. Those that were considered positive in the past might be presently be considered negative, leading to polarity shifts[84]. In reference to subjectivity, different people may have different perspectives on what is positive or negative, resulting in discrepancies in sentiment analysis [86]; [76]. Sentiments can be cultural and regional-specific and what is considered positive in one culture or region might be perceived differently in another, leading to polarity shifts [106]. Sentiments can change based on the current news, events, or trends.

For instance, a product that was well-received might suddenly receive negative sentiments due to a recall or controversy [102]. The tone and style of the author can influence how a text is interpreted. For example, a sarcastic or critical tone may make a generally positive statement appear negative [104]. In reference to review frauds, there can be fake or fraudulent online reviews and ratings that are intentionally designed to mislead sentiment analysis systems, resulting into polarity shifts [107]. Pragmatic aspects of language, such as implicates and presuppositions, can lead to polarity shifts. The meaning can depend not just on the words used but on the speaker's intended meaning [108]. Lastly, in domain-specific sentiment analysis, the absence of domain-specific knowledge may lead to polarity shifts, as certain terms or expressions may have unique connotations in a particular domain [109].

Sentence contrast occurs when the neighbouring sentences' polarities diverge from their intended direction. Explicit contrast is the term used for this change in polarity. If the contrast is taken into consideration, the foremost half of the review's positive sentiment will be reversed by the presence of a contrast, but the entire review will take on a negative emotion[102]. Connecting one phrase, sentence, or paragraph to another one expresses contradictions which fall under the intra-sentences, extra-sentences and extra-paragraph categories [76]. When there are inconsistencies or conflicting sentiment patterns in the reviews, a form of polarity shift occurs. Such a change can take place in lengthy evaluations if the reviewer presents opposing perspectives about many aspects of the good or service [110]. In this instance, the reviewer's opinion on one side of the product differs from their opinion of the product as a whole.

The factors discussed above contribute to the sentiment classification accuracy because they reverse the polarity of a sentence from what was initially intended. The performance of sentiment classification model is significantly impacted since not all polarity shifters have been incorporated into the existing sentiment analysis models used in detecting and handling the polarity shift problem. One of the major problems with sentiment analysis is the shift of polarity. The main causes of polarity shift are referred to as “polarity shifters," a set of linguistic contextual factors that can change a word's prior polarity in one of the three ways i.e. rise, decrease, or neutral.

According to Madhuli and Rahuli [13], there is need to consider all the factors that are responsible for polarity shifts such as contrast transition, sentiment inconsistency, exception structure and negation. The performance and accuracy of sentiment classification models can be enhanced by the automatic detection of such factors. If not detected and handled, a number of contextual shifters can have an impact on the polarity of the attitude [100].

Negation words like not, neither and other similar phrases can change the polarity of a sentence, thus making them crucial for sentiment analysis [102]. Positive expressions include "this movie is good" and negative ones include "the movie is not good". Most of the time, certain approaches eliminate negation terms because they are on stop-word lists or are implicitly disregarded[81]. The polarity shift of a sentence is determined by the two essential components of negation, cue and scope.

The covering of negation's effects on the sentence is its scope. According to Nikhil [103], classification of negation cue is listed into three categories i.e. linguistic negations, structural and valence shifters. Negation in linguistics can either be explicit or implicit. While implicit negation, such phrases as rarely, avoid, etc., is indirect, explicit negation expresses itself clearly. Contrast is the factor that causes structural negation occur most often, where contradiction can be expressed with the connecting sentence, paragraph, etc. using connectors like but, however, etc. [6]. Valance shifters belong to the categories of diminisher and intensifier; a diminisher reduces the opinion intensity or polarity of the word to which it is linked, for example, less few, small. An intensifier raises the opinion intensity or polarity of the term. While intensifiers, such as badly, epic, absolute, etc., strengthen the opinion term with which they are linked. A given polarity's intensity will be increased by intensifiers in either the positive or negative extreme. To be considered an intensifier, a shifter must have an overall impact on the polarity; it could lead to a rise in intensity, a graduation in the same pole, or a decrease in another pole [81]. Similarly to the way a shifter must generate a final effect in this case, an improvement in the strength or improvement on the polarity scale in similar pole in order to be classified as a diminisher is critical. In specific, the positive or negative pole diminishers decrease a specific polarity [81].

A number of implicit negations have been discussed in linguistic and are relevant to sentiment analysis. Firstly, lexical implicit negations that often carry a negative meaning without explicitly using negators are types of implicit negations. They use antonyms with negative connotations where negative counterparts are used for example “bad” instead of “not good”. On the other hand negative prefixes and suffixes are used to express lexical implicit negations. Words like “un”,”in”,”im” and “non” are used to express the implicit negation. Nouns with negative implications and verbs with connotation are also used to express lexical implicit negations in sentiment analysis. Secondly, Morphosyntactic implicit negations that involve grammatical structures are used to express implicit negation. These include negation quantifiers and determiners, restrictive adverbs and positive adjectives or adverbs which often imply a negative outcome. Lastly, implicit negation can be influenced by semantic and pragmatic facts that rely heavily on the meaning and context of the sentence. These include rhetorical questions with negative

implications, irony and sarcasm which convey a negative sentiment opposite of what is meant. Implications through contrast, presupposition and modal verbs with negative implication are semantics used to express implicit negation in sentiment analysis. In this study, both lexical implicit negation and semantic negations that included contrast were the main focus. This allowed the researcher to narrow the scope of the study.

2.5 Review of the Existing Sentiment Analysis Models for Handling Polarity Shift

This section describes a comprehensive review of the existing sentiment analysis models used in detecting and handling the polarity shift problem. The literature comprised of both the deep learning and machine learning models used to handle the polarity shift problem. A systematic literature review methodology was adopted to ensure quality of the review process. The review took a narrow focus on investigating the architecture across the different models and their common challenges, common polarity shifters across the models and the nature and dimensionality of datasets across them. It also focused on investigating the trends on the feature selection methods as well as the key metrics for assessing the effectiveness of these models.

The review was guided by the research questions formulated as follows:

RQ1: What implicit sentiment inversion factors affect polarity shifts in sentiment classification within written texts?

The systematic literature review adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) technique to ensure quality literature search. Figure 2.6 shows how the analysis process was conducted.

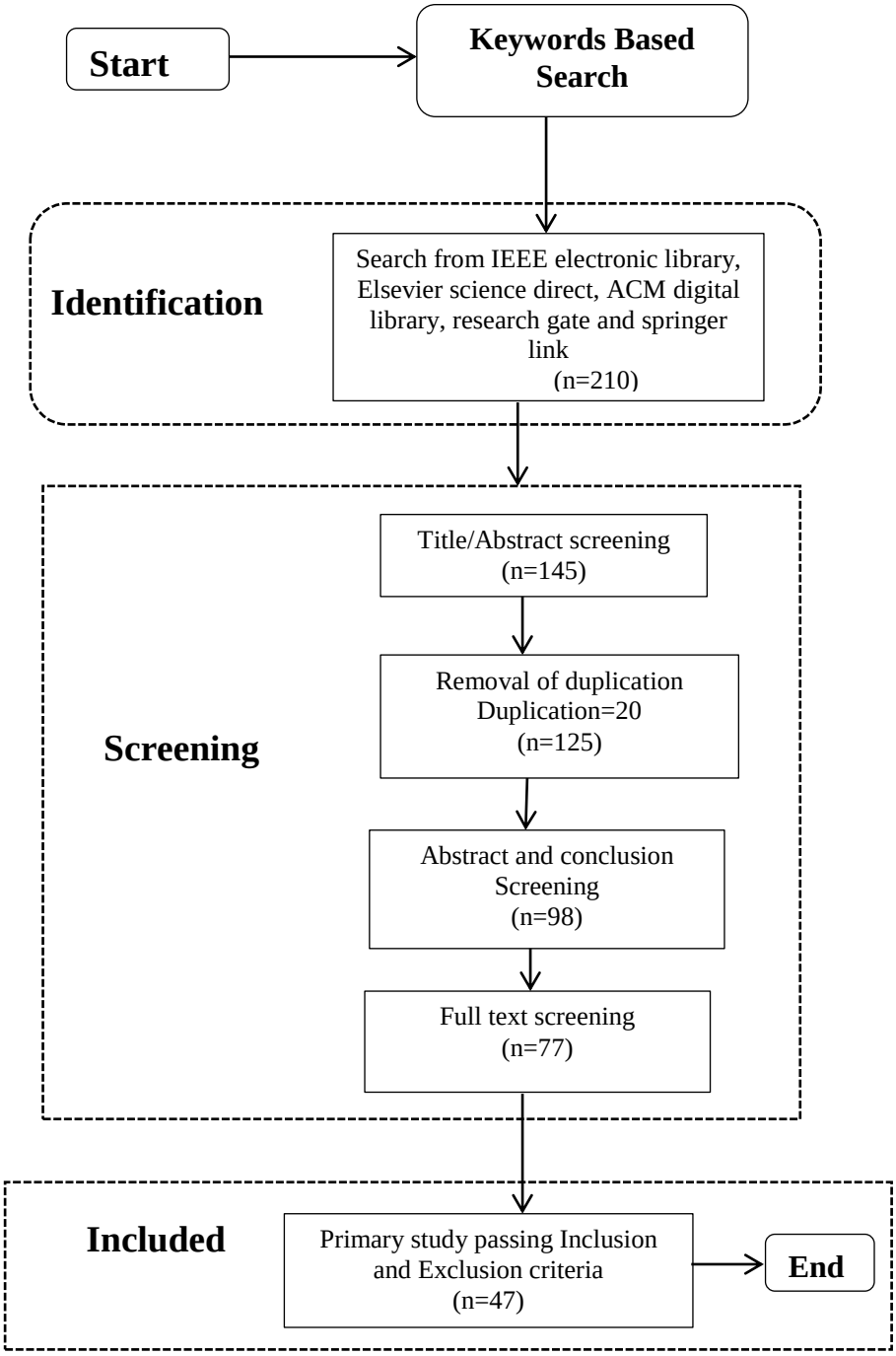


Figure 2.6: PRISMA Flow Chart for the Articles Selection[111]

The technique comprised of three stages namely; identification, screening, and inclusion as shown in figure 2.6. The review was limited to journals and conference proceedings written in English from the scientific digital databases published between 2018 and 2024. The search utilized the scientific databases namely IEEE electronic library, Elsevier science direct, ACM digital library, research gate and springer Link. The review resulted to a total of 47 peer-reviewed journals that were eligible for analysis. The 47 peer–reviewed journals were justified by the quality

assessment and selection criteria that ensured the articles were of greater relevance on research question for the review.

2.5.1 Sentiment Analysis Models used in Detecting and Handling the Polarity Shift Problem

Sentiment analysis aims to classify words as positive, negative, or neutral by identifying those conveying sentiment in unstructured text inputs. Among the main problems with sentiment analysis is the polarity shift. According to research, the accuracy of sentiment analysis is improved by polarity shift detection [81].Researchers have proposed models for addressing the problem of polarity shifts in sentiment analysis [112]. Machine learning and deep learning models are the two main types of sentiment analysis models currently used in detecting polarity shifts. The subsequent sections presents the summary in tables of the sentiment analysis models, the key polarity shifters, the inputs datasets used with the sentiment analysis models, the feature selection techniques and the key metrics used the models. A number of sentiment analysis models used in the detection and polarity classification is discussed highlighting their limitations.

LSTM based deep neural network approach [81] using Bidirectional LSTM was proposed to recognize negation cues in sentences. It was trained using word level features. The approach only handled negation polarity shifters.

Attention-based Bidirectional CNN-RNN deep model [113] that utilized two independent bidirectional LSTM and GRU layers, to extract both past and future was proposed. The authors applied convolution and pooling mechanisms that reduced dimensionality of features and extract position invariant local features it. The main limitation of the model is that it focused on document level polarity detection.

Supervised Contrastive Pre-Training to learn implicit negation[114] model was proposed. The model aligns the representation of implicit sentiment expressions to those with the same sentiment label. The pre training process then leads to capture of both implicit and explicit sentiment. The main limitation is that the sentiment model requires knowledge-enhanced methods to improve the accuracy when detecting the implicit negation.

Hierarchical multi-task learning approach [115] was proposed that incorporates the information about negation by using cascading and hierarchical neural architecture. The model was trained with negations as an auxiliary task. The model experimented with negation, however, there is need to include other factors like linguistic and paralinguistic phenomena, implicit negations, which also affect sentiment classification

Recursive neural network based on sequence labelling model [116] that utilized cue word and its target scope was proposed. It identified negations using the global dependency structure and learns a high-level representation for each word in the context of each sentence. The representation from the recursive neural network is then fed into the Conditional Random Fields (CRF) layer, which collaborates to jointly decode the labels for the entire sentence. However, the constituency tree and dependency tree features were not integrated by the model to enable implicit negation scope detection.

Reinforcement learning framework [117] that uses dynamic learning framework that detects negations was proposed. It used document level labels as input and learns negation policy. The model aimed to replicate human perceptions; however, the model handled a limited number of polarity shifters. It is key to include implicit negations inorder to improve accuracy.

Opinion Mining based complex polarity shift pattern [102] was proposed to determine which machine learning model best fits the dataset. The model handled the following polarity shifters; contrast transitions, intensifiers, diminishers, and negatives were all handled. Word sense disambiguation caused by implicit negations polarity shifters that can enhance the sentiment classification process were not fully addressed by the model.

Stacking Ensemble Method [118] was proposed to handle preprocessing, polarity shift detection and elimination, and sentiment classification. The stacking ensemble method combined four modules. This model did not handle complex polarity shift structures like implicit negations, sentence inconsistencies and contrasts.

Semi-supervised domain adaptive dual sentiment analysis. [119] that converts the reviews to its opposite review was proposed. The Bayesian classifier was used to train both original and reversed reviews. Probabilities of the tuples or the specified class are used by the classifier. The reviews are put to the test using later dual prediction. LSTM handled the classification while CNN handled feature extraction. This model was trained to handle only domain specific polarity shifter.

Negation detection based on deep learning approach (BERT) [120] model was proposed. This model looks for negation in Spanish-language product reviews using a transformer-based approach. It performs negation detection using a BERT-based model and transfer learning techniques. However, the model did not perform the sentiment analysis tasks. It only detected the negation in the sentences.

Transfer learning approach to handle negations [121] model was proposed. The model used Neg BERT architecture which employed a series of tokenized and encoded sentence tokens as the input. For all three datasets utilized, cue detection and scope resolution were carried out; one dataset was used for training, and the other two for testing. However, the model may require a bigger dataset in order to extract the maximum generalizability from the architecture.

Negation scope detection based on recurrent neural networks technique [122] was proposed. In this, sequence labelling was implemented in the model, whereby each word in a sentence was designated as either inside (I) or outside (O) of the scope. The RNN model classified every word in a sentence as belonging to the scope or not. The RNN layer then receives as inputs the word embedding model that was produced as well as the sentence's index representation. It is done to predict each label's probability. Lastly, the labels that each word in the sentence corresponds to are the model's output. The main limitation is that the model only performed explicit negation detection. There is need to perform sentiment analysis.

Negation and assertion detection based on attention-based deep learning system [123] that identified assertions and negations. The model employed word embeddings and attention-based BiLSTM networks. This approach focused on scope resolution in order to solve NLP issues. However, the model was domain-specific and handled the explicit negation only.

The most prevalent sentiment analysis models shared the deep learning architecture. This is due to the fact that feature engineering doesn't have to take as long because deep learning architecture automatically represents text features. Additionally, it was pointed out that deep learning architectures use word embedding to represent text using dense vector representations; as a result, automated feature input replaces the need for manually created features. On the other hand, some models have adopted rule based approach because of its fact that it does not need tuning of the parameter.

It is also applicable where linguistic complexity does not appear after the sentence. However, rule based approaches are domain dependent and are mostly applied where there is predefined features that are static in nature. It is noted that though machine learning based models have been used in handling polarity shift, it is only applicable where sufficiently large labelled data for model fitting is available.

Most of the machine learning models for handling polarity shift suffers from lacks of generalizability. It is worth noting that sentiment analysis is a sequence tasks therefore Hidden Markov Model (HMM) and Conditional Random Field (CRF) are the two algorithms that have been used. However, since HMM and CRF are both Markov chain-based models, they are unable to handle long sequential dependency. More so, it was noted that the efficiency of hybrid classifiers is higher than that of any single classifier, particularly in situations where there is insufficient training data.

Figure 2.7 summarizes the sentiment classification models used in handling polarity shift based on their popularity. As shown below, deep learning models have received a lot of attention over traditional machine learning and hybrid or ensemble model.

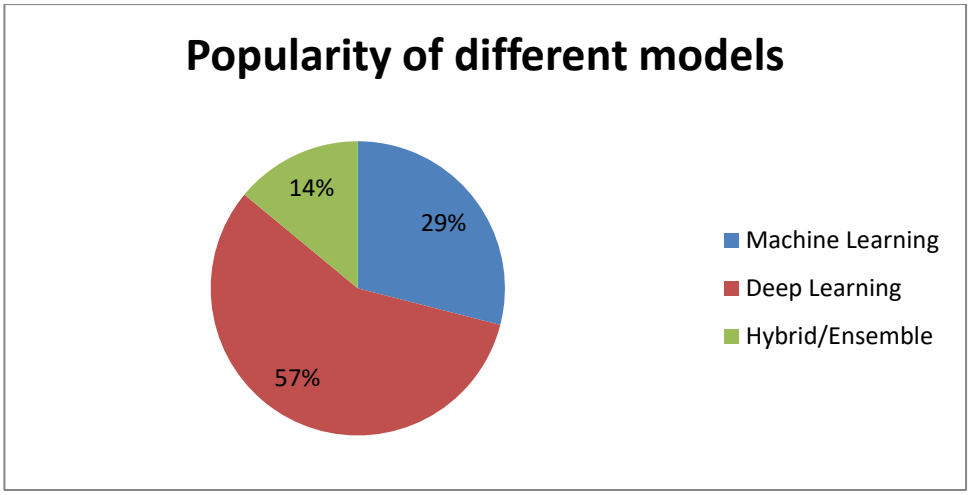


Figure 2.7: Research Trends on the Sentiment Analysis Models [124]

The commonly applied traditional machine learning models include Support Vector Machines, Naive Bayes, and Logistic Regression. These models demonstrate robustness in certain contexts. Furthermore, deep learning models, notably LSTM, GRU, and BERT, are also deployed. The latter offers advanced capabilities in capturing nuanced sentiment variations and subsequently improved polarity shift management. Hybrid models and ensemble techniques, combining the strengths of different approaches, have also been applied. These models exhibit a rich diversity that contributes to enhanced efficacy in managing polarity shifts. The efficacy of sentiment analysis models in polarity shift management is evident through their ability to adapt to changing sentiments across different contexts.

2.5.2 Polarity Shifters Commonly Detected by the Existing Sentiment Analysis Models

This section discusses the common polarity shifters and their effect in sentiment classification. Research show that the traditional machine learning models such as Support Vector Machines, Naive Bayes, and Logistic Regression exhibit a certain level of dynamism, particularly in capturing explicit shifts in sentiment. However, deep learning models, such as LSTM and BERT, offer a more dynamic interpretation of sentiment shifts, particularly in nuanced and context-dependent situations. The incorporation of contextual representations and sophisticated embeddings increases the dynamic of these models, allowing them to detect small modifications in sentiment expressions associated with polarity shifts. The systematic review reveals

that among various factors influencing polarity shift, negation emerges as the most significant polarity shifter.

Sentiment classification models commonly encounter instances where the presence of negation alters the polarity of sentiment expressions. Explicit negation, such as the use of "not" or "but," is effectively handled by most models. However, there is few evidence of the effectiveness of these models in managing implicit negation. The commonality of negation as a polarity shifter highlights its importance in influencing sentiment dynamics, necessitating models that can discern and appropriately adjust sentiments in the presence of negating language. A summary of the polarity shifters commonly detected by sentiment classification is discussed.

Negation plays a significant role in written text. Negations are inherently dynamic in nature because of their association with linguistic complexity. They can be explicit or implicit, and their use is highly domain-specific [81]; [115]. Explicit negation with a larger scope directly impacted by compositional phenomena, is the most common of negation type. It is notable that in order to enhance the precision of sentiment classification, the classifier must possess the ability to recognize implicit negation and separate the impact of its extent on the ultimate polarity of a given text [116]. Negation and speculation scope detection is a sequence labelling task. Therefore, there is need of employing labelling scheme for each token inorder to detect the implicit negation. The presence of negatives makes it challenging to interpret the narrative's meaning. Negation use is domain specific therefore there is need to use a dynamic learning to detect negations [117]. Typical polarity shifters include negativity, contrast, and sentiment inconsistency [118]. The use of negation words both implicit and explicit and contrasts causes machine learning algorithms to classify the text as positive or negative when it isn't. The works of [102] identifies negations, intensifiers, diminishers and contrast transitions as the main polarity shifters. It is noted that complex linguistic structures like implicit negation can change or reverse the text's sentiment polarity. Domain specific shifters [119] are key in sentiment analysis. Classifier trained in one domain fails to perform well in another domain. This will require the model to be trained once more for the other domain. The works of [120]; [77]notes that a negation trigger affects a set of tokens,

known as scope, and negation detection is divided into two subsets: scope resolution and trigger detection.

The review highlights a clear trend toward improving sentiment analysis models' dynamism in order to handle polarity shifts more effectively. The models' ability to adjust to changing sentiments in a variety of situations reflects this dynamism. Traditional machine learning models such as Support Vector Machines, Naive Bayes, and Logistic Regression exhibit a certain level of dynamism, particularly in capturing explicit shifts in sentiment. However, the deep learning models, including LSTM and BERT are good in providing a more dynamic understanding of sentiment changes, especially in nuanced and context-dependent scenarios. These models are made more dynamic by the use of contextual representations and advanced embeddings, which allow them to detect subtle variations in sentiment expressions linked to polarity shifts. The systematic review reveals that among various factors influencing polarity shift, negation emerges as the most significant polarity shifter.

2.5.3 The Input Datasets used with the Sentiment Analysis Models for Detecting Polarity Shift Problem

Although most of input datasets into these models were structured, there were a number of datasets that were unstructured. Most of the input datasets into these models were text and consisted of thousands of reviews. It is clear that some researchers have used Simon Fraser University (SFU) review corpus Negation Speculation and Conan Doyle-neg data sets [81]. This is because they are the text dataset that is available for study, particularly in negation handling, and they are annotated with negation cues. The 400 narrative texts in the Simon Fraser University (SFU) review corpus comprises of the consumer product, movie, and book domains.

The review emphasizes the significance of domain-specific considerations in shaping effective sentiment analysis models for managing polarity shifts. The nature and characteristics of the dataset play a pivotal role in the adaptability and performance of sentiment analysis models. Incorporating domain-specific datasets that capture the nuances of sentiment expressions within industries or themes is becoming more and more popular, according to the review. These datasets are made to account for the subtleties and distinctive language characteristics connected to specific domains, allowing sentiment analysis algorithms to behave in a domain-aware manner. By

acknowledging domain specificity, the models become more dynamic and are better equipped to handle polarity shifts in contextually different domains. One important aspect of sentiment analysis models that affects the extent to which they handle polarity shifts is the labelling process. The systematic review reveals a shift towards employing more meticulous and context-aware labelling processes. There is a growing emphasis on ensuring that labels accurately represent the sentiment expressed within the given context, particularly considering the dynamic nature of polarity shifts.

2.5.4 Feature Selection Techniques applied in the Sentiment Analysis Models used for Detecting and Handling the Polarity Shift Problem

The review indicates a growing preference for sophisticated feature selection techniques that enhance model adaptability. Recursive Feature Elimination stands out prominently, demonstrating effectiveness in preserving relevant features crucial for capturing nuanced sentiment changes. Additionally, the emergence of Term Frequency-Inverse Document Frequency as a popular filter-based method suggests an inclination towards techniques that consider both term frequency and document specificity, optimizing the selection of informative features [24]. The review highlights a shift towards more intricate hybrid feature selection methods, with Random Forest gaining prominence for its ability to capture complex relationships within the data. Firefly (FF) and Levy flights (FFL) has been used sentiment analysis [125]. The main challenge with this technique is the disparity of light intensity and the brightness formulation. Entropy-based metric and an evolutionary algorithm-Hybrid model[126] has been utilized. However, the technique does not control the feature vectorization step. It needs to be expanded to more datasets from different domains. SMOTE and Information Gain has been used with the machine learning algorithms [127]. GAWA- Included Wrapper Approaches (WA) has been used to select the premier features and the Genetic Algorithm (GA) to reduce the size of the premier features[128]. The technique can be examined with multiple datasets from variety of sources in order to select the best features with various categories of syntactic and stylistic features. Chi square-based feature selection[129] has been utilized to enhance the technique performance.

It is observed that the view of feature representation in text data has made significant advances toward embedding techniques. The most recent advancement in representing documents as vectors is the ability to embed words, characters, sentences, and paragraphs. Word embeddings are widely used by deep learning-based models to represent text in vector representations. Contextual embedding is the recent advancement in the feature extraction based on context and has shown state-of-the-art performance in sentiment analysis. BERT models and its variants like RoBERTa and DistilBERT have been used for contextual representation. Therefore, this acts as an automatic input features instead of using hand-crafted features. It is noted that feature selection process involves choosing essential parameters; hence this improves the computation of a classifier method. It is noted that hybrid feature selection has been commonly used in the sentiment analysis models. This is due to the fact that hybrid models obtain the most optimal features while reducing the original feature space. It is important to consider the characteristics of the dataset when selecting a feature selection technique.

2.5.5 Metrics used to assess the effectiveness of the Sentiment Classification Models used for Detecting the Polarity Shifts

The systematic review delves into the critical aspect of performance evaluation metrics. This offers insights into the methodologies employed to assess the efficacy of sentiment classification models. The selection of appropriate metrics is paramount in gauging the models' effectiveness in capturing nuanced sentiment variations across diverse contexts. The review identifies several commonly used evaluation metrics, showcasing their relevance in assessing different aspects of sentiment analysis models. Accuracy, precision, recall, and F1-score are identified as the traditional metrics that measure the overall correctness and balance of sentiment classifications. An extensive evaluation of the performance metrics used to evaluate the models used in sentiment classification is done. In the works on Negation and Speculation Detection ML approach [130], the authors utilized F1 measure, Geometric(G)-mean, percentage of correct relaxed scopes (PCRS), PCR. Multi-task approach [115] works utilized accuracy. LSTM based deep neural network model [81] utilized F1-Score to evaluate their model. Recursive neural network sequence labelling model [116] utilized F1 score and Percentage of Correct Scopes (PCS) for detection of negation scope model evaluation. The opinion mining based complex

polarity shift pattern [102] model utilized confusion matrix, accuracy, precision, F1 Score, and Recall. Stacking Ensemble Method[118] proposed for negation detection utilized recall, precision, and accuracy. Deep learning approach for negation detection [120] model utilized F1 score for the cue detection and scope resolution detection evaluation. Attention-based deep learning system for negation and assertion detection [123] utilized precision, recall and F1 scores on negation was done. Opinion mining model for handling polarity shift [77], utilized 10-fold cross validation for the model evaluation.

The performance of the models used to solve the polarity shift problem has traditionally been assessed using three metrics: accuracy, cross validation, and F1 score. The literature review suggests that the most popular metrics for assessing these models' performance are accuracy, F1 score, and cross validation. There is need to prove the evaluation metrics that would be the most suitable in some application areas. Precision can be useful when the cost of false positives is high i.e. misclassifying a neural or negative sentiment as positive. For example, in a sentiment analysis model for product recommendation, precision would be the most suitable metric as it would avoid recommending a product that customer ultimately dislikes.

In the recent development techniques like LIME,SHAP, integrated gradients and attention have been used to measure the models' interpretability and explainability[19]. LIME (Local Interpretable Model-agnostic Explanations) is a technique used to explain the predictions of any machine learning model, regardless of its complexity. It explains individual predictions rather than providing a global understanding of the model's behaviour [131]. SHAP (SHapley Additive exPlanations) is a technique used to explain the output of machine learning models. It provides a way to understand the contribution of each feature to individual predictions [132]. SHAP quantifies the importance of each feature by considering all possible combinations of features. Integrated gradients can help explain the reasoning behind predictions made by deep learning models in tasks [133]. Integrated Gradients make it possible to examine the inputs of a deep learning model on their importance for the output.

2.5.6 Discussions on the Literature Review Analysis

The culmination of the critical review analysis on the sentiment classification models used in handling polarity shift reveals complex trends and insights that highlight limitations and advancements in this field. This not only assists in emphasizing the important patterns, themes, and theories from the literature review, but it also offers insightful information and implications for future research. In the subsequent sections, a critical evaluation and summary on the existing literature in regards sentiment classification models used in handling polarity shift and feature selection and extraction techniques used with these models. Lastly the datasets used and metrics for evaluating the performance of these models is discussed. Firstly, the review is organized based on topics, themes and key theories. This approach is in line with the works of Snyder (2019) on literature review as a research methodology.

2.5.6.1 Sentiments Classification Models' Efficacy and Future Research

The review reveals a diverse landscape of sentiment analysis models employed for polarity shift management. Different models showcase varying approaches to handling the dynamic nature of sentiment polarity. The findings underscore the importance of model diversity and highlight the nuanced effectiveness of specific techniques in managing polarity shifts within sentiment analysis. Exploring the impact of temporal aspects on sentiment dynamics and developing models that can adapt to evolving language trends presents an exciting avenue for future studies. Future research endeavors should focus on refining existing models, exploring innovative hybrid approaches, and developing models that can handle diverse and domain-specific datasets. For example, addressing the challenges associated with implicit negation represents a critical area for exploration, requiring models that can discern and handle subtle shifts in sentiment expression.

2.5.6.2 Common Polarity Shifters and the Dynamism in Handling Polarity Shift

The systematic review reveals that among various factors influencing polarity shift, negation emerges as the most significant polarity shifter. Sentiment analysis models commonly encounter instances where the presence of negation alters the polarity of sentiment expressions. Explicit negation, such as the use of "not" or "but," is effectively handled by most models. However, there is limited evidence of the effectiveness of these models in managing implicit negation, indicating a potential

area for improvement. The commonality of negation as a polarity shifter highlights its importance in influencing sentiment dynamics, necessitating models that can discern and appropriately adjust sentiments in the presence of negating language. Addressing the complexities of negation in sentiment analysis models is crucial for enhancing their overall dynamism and effectiveness in managing polarity shifts.

The review underscores a pronounced trend toward enhancing the dynamism of sentiment analysis models in effectively managing polarity shifts. The dynamism is reflected in models' adaptability to changing sentiments across diverse contexts. Traditional machine learning models such as Support Vector Machines, Naive Bayes, and Logistic Regression exhibit a certain level of dynamism, particularly in capturing explicit shifts in sentiment. However, the deep learning models, including LSTM and BERT that excel in providing a more dynamic understanding of sentiment changes, especially in nuanced and context-dependent scenarios.

2.5.6.3 Trends in Feature Selection and Feature Extraction Techniques used with Sentiment Classification Models

This review unveils notable trends in feature selection and feature extraction techniques employed in sentiment classification models used in managing polarity shifts. The trends indicate a growing preference for sophisticated feature selection techniques that enhance model adaptability. Recursive Feature Elimination stands out prominently, demonstrating effectiveness in preserving relevant features crucial for capturing nuanced sentiment changes. Additionally, the Term Frequency-Inverse Document Frequency, a popular filter-based method suggests an inclination towards techniques that consider both term frequency and document specificity, optimizing the selection of informative features. The review highlights a shift towards more intricate hybrid feature selection methods, with Random Forest gaining prominence for its ability to capture complex relationships within the data.

The review on the feature extraction indicates a shift towards leveraging advanced word embeddings and contextual representations. Word2Vec remains a consistent trend as an effective word embedding technique, showcasing its continued relevance in capturing semantic relationships. BERT, representing contextual embeddings, emerges as a dominant trend, highlighting the increasing importance of contextual information in understanding sentiment nuances. The use of such sophisticated

embeddings signifies a trend towards more nuanced and context-aware sentiment analysis models capable of addressing polarity shifts. These identified trends collectively suggest a paradigm shift towards more intricate, context-aware and adaptable sentiment analysis models for managing polarity shifts. The integration of advanced feature selection and extraction techniques reflects models that can effectively capture the dynamic nature of sentiment in evolving language contexts.

2.5.6.4 Domain-Specific Considerations for the Datasets and the Labelling Process

The review emphasizes the significance of domain-specific considerations in shaping effective sentiment analysis models for managing polarity shifts. The nature and characteristics of the dataset play a pivotal role in the adaptability and performance of sentiment analysis models. The review identifies a growing trend toward the utilization of domain-specific datasets that capture the intricacies of sentiment expressions within industries or topics. These datasets are designed to accommodate the nuances and unique linguistic features associated with specific domains, enabling sentiment analysis models to exhibit domain-aware behaviour. In the realm of sentiment analysis models, the labelling process is a critical factor influencing their effectiveness in managing polarity shifts. The systematic review reveals a shift towards employing more meticulous and context-aware labelling processes. There is a growing emphasis on ensuring that labels accurately represent the sentiment expressed within the given context, particularly considering the dynamic nature of polarity shifts.

2.5.6.5 Performance Evaluation Metrics

The systematic review delves into the critical aspect of the performance evaluation metrics, offering insights into the methodologies employed to assess the efficacy of sentiment analysis models in managing polarity shifts. The selection of appropriate metrics is paramount in gauging the models' effectiveness in capturing nuanced sentiment variations across diverse contexts. The review identifies several commonly used evaluation metrics, showcasing their relevance in assessing different aspects of sentiment analysis models. Accuracy, precision, recall, and F1-score are identified as the traditional metrics that measure the overall correctness and balance of sentiment classifications. However, the review emphasizes the importance of metrics tailored

to polarity shift management. Cohen’s Kappa and Matthew's Correlation Coefficient are highlighted for their ability to account for imbalanced datasets and measure the models' agreement with ground truth labels, providing a more comprehensive evaluation. Acknowledging the influence of domain specificity, the review notes the adoption of domain-centric evaluation metrics.

It is clear from literature that there is critical need to improve sentiment analysis models, particularly in contexts where implicit negation introduces significant polarity shifts. While sentiment analysis has made considerable progress, current models still struggle with accurately identifying sentiment when subtle implicit negations or contextual nuances are present. This issue is especially prominent in domains such as mobile phone reviews on e-commerce platforms like Amazon, where implicit negations such as "could have been better than “or” not as great as expected" are frequent and convey crucial sentiment information that current models fail to fully capture effectively. Polarity shifts, arising mainly from the implicit negation, remain a major challenge, leading to misclassifications that undermine the accuracy and reliability of sentiment analysis. These misclassifications have far-reaching implications for businesses, consumers, and automated systems that rely on sentiment data to make decisions.

The review reveals a trend towards developing sentiment analysis models that exhibit a higher degree of generalizability, enabling them to adapt to varying linguistic patterns, contexts, and domains. The integration of domain-specific considerations, robust feature selection techniques, and diverse datasets contribute to the enhanced generalizability of sentiment analysis models, allowing them to navigate and manage polarity shifts in a broader spectrum of applications. The systematic review recognizes the evolving landscape of sentiment analysis The potential impact of this work is significant, not only for enhancing the accuracy of sentiment classification but also for improving real-world applications, such as consumer feedback analysis, market sentiment prediction, and recommendation systems.

2.6 Implications of the Key Findings in the Experimental Design Process

The review reveals a trend towards developing sentiment analysis models that exhibit a higher degree of generalizability, enabling them to adapt to varying linguistic patterns, contexts, and domains. Generalizability refers to a model's ability to perform well across diverse datasets and real-world scenarios. The integration of

domain-specific considerations, robust feature selection techniques, and diverse datasets contribute to the enhanced generalizability of sentiment analysis models, allowing them to navigate and manage polarity shifts in a broader spectrum of applications.

The systematic review recognizes the evolving landscape of sentiment analysis models with a distinct focus on embedding and representation learning techniques. There is a noteworthy trend in leveraging transfer learning as a mechanism to enhance generalizability. Pre-trained models, especially those based on deep learning architectures like BERT, are fine-tuned for sentiment analysis tasks. This transfer learning approach allows models to leverage knowledge gained from large, diverse datasets, improving their adaptability to specific sentiment analysis tasks and boosting generalizability across various domains. Despite advancements, challenges exist in achieving optimal generalizability. Implicit negation, domain-specific linguistic nuances, and dynamic shifts in sentiment expressions are identified as the challenges that accurate sentiment classification.

The key findings reveal the evolving research in the sentiment analysis and their effectiveness in the detection and handling of polarity shift with the aim of improving the sentiment classification. The review highlights recent advancement in the sentiment classification models, feature selection and the data preprocessing techniques. These findings provide a clear direction for the subsequent empirical analysis which addresses objective number two. The integration of literature findings into the methodology not only advances knowledge within this field but also contributes to the progression of advanced research methodologies.

2.7 Conceptual Framework

The conceptual framework defines the conceptual elements, relationship between dependent and independent variables as well as the operationalization of the variables [92]. In this research work, the conceptual framework has formulated the independent variables, moderating variables and dependent variables. The independent variable is the sentiment analysis model. The newly developed hybrid model for detecting the polarity shift due to implicit negation is used to construct the independent variables. The feature extraction, feature selection and data augmentation techniques are used as the moderating variables. Moderating variables

provides a mechanism of how the independent variables influence the dependent variables. In this case, the choice of feature extraction, feature selection and data augmentation technique used on the polarity shift handling models will give a varying score of the dependent variable i.e. the evaluation metrics of the sentiment classification model.

2.7.1 Independent Variables

The hybrid deep learning model was identified as the independent variable. The deep hybrid model was used in the validation process in order to evaluate its level of effectiveness against the existing best performing models, in a variety of mobile phone datasets. The hybrid model's accuracy was compared against four sentiment classification models and their hybrids. The models includes: two traditional machine learning (support vector machine and naïve bayes) and two deep learning models (LSTM and BERT).

2.7.2 Moderating Variables

Moderating variables are variables that act between the independent variables and the dependent variables and change the strength of their relationships by influencing the predictor and the outcome [92]. They provide a potential explanation or mechanism for how the independent variable influences the dependent variable. In this research study, the feature selection and feature extraction as well as the data augmentation techniques are adopted as the moderating variables. The effectiveness of the sentiment classification model is defined by the type of feature selection, feature extraction and the data augmentation techniques used.

2.7.3 Dependent Variables

In this research, the dependent variables are as a resultant of the hybrid deep learning model's ability to detect and classify implicit negation effectively. The effectiveness of the hybrid model was evaluated using the Accuracy, Mathew's correlation Coefficient and Cohens Kappa. Cohen's Kappa and Matthew's Correlation Coefficient effectively measure the models' agreement with ground truth labels, providing a more comprehensive evaluation. During the validation process, the three metrics are aggregated into a single metric called the average performance score and are used to evaluate the proposed model against the existing best performing ones in a variety of datasets. An additional of the metrics namely; Mean Squared Error,

Kruskal Wallis H statistics, ANOVA and attention heat maps are used to ensure the generalizability of the newly developed model. Figure 2.8 illustrates the relationship between the independent variables, moderating variables and the dependent variables in this research's conceptual model.

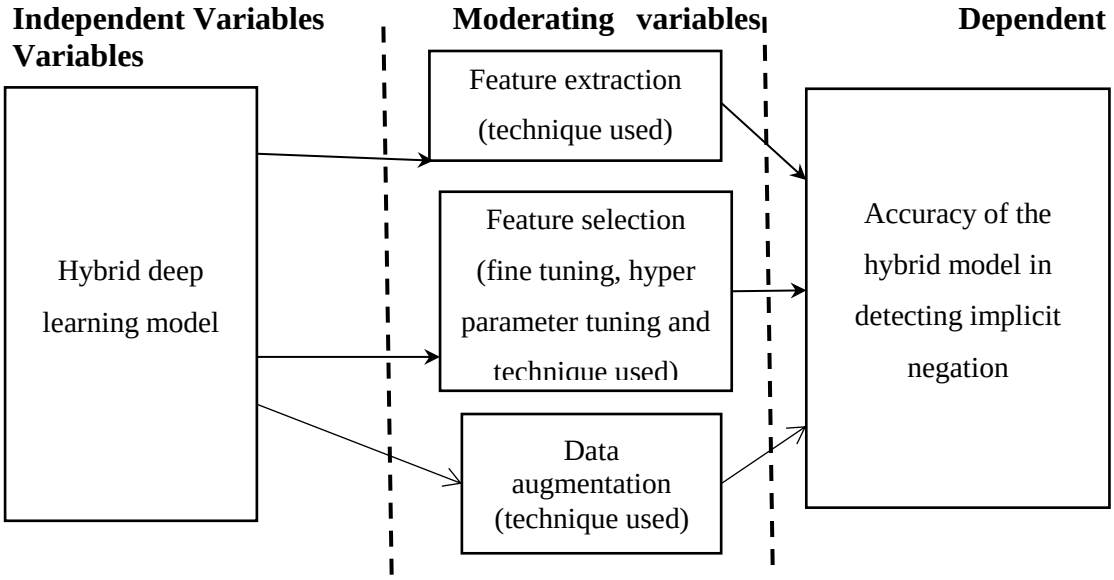


Figure 2.8: Relationship between independent, moderating and dependent variables

2.8 Summary

The chapter presents a review of the sentiment analysis models used for detecting the polarity shift problem. The review findings show that polarity shift is the main challenge with the sentiment analysis models. It is clear that current sentiment analysis models for detecting the polarity shift problem have not been effective in handling all polarity shifters, leading to poor performance of the classification models. The review also presents the input datasets used with the sentiment analysis models as well as the feature selection, feature extraction and data augmentation techniques and the metrics for evaluating the models' effectiveness.

A systematic literature conducted establishes that negation is the main polarity shifter in sentiment analysis models. It is evident from the review that there was no single model that could universally detect the polarity shift problem. At the same time, most of these models detect relatively few polarity shifters. The review underscores a pronounced trend towards enhancing the dynamism of sentiment

analysis models in effective detection of polarity shifts. The dynamism is reflected in models' adaptability to changing sentiments across diverse contexts. Traditional machine learning models such as Support Vector Machines, Naive Bayes, and Logistic Regression exhibit a certain level of dynamism, particularly in capturing explicit shifts in sentiment. However, the deep learning models, including LSTM and BERT excel in providing a more dynamic understanding of sentiment changes, especially in nuanced and context-dependent scenarios. The integration of advanced embeddings and contextual representations contributes to the dynamism of these models, enabling them to capture subtle variations in sentiment expressions associated with polarity shifts.

The systematic review reveals that among various factors influencing polarity shift, negation emerges as the most significant polarity shifter. Explicit negation is effectively handled by most models. However, there is limited evidence of the effectiveness of these models in managing implicit negation, indicating a potential area for improvement. The commonality of negation as a polarity shifter highlights its importance in influencing sentiment dynamics, necessitating models that can discern and appropriately adjust sentiments in the presence of negating language. Addressing the complexities of negation in sentiment analysis models is crucial for enhancing their overall dynamism and effectiveness in detecting polarity shifts.

The review shows that deep learning models, particularly transformer-based architectures like BERT and GPT, have significantly improved sentiment analysis by capturing contextual nuances. However, these models still struggle with polarity shifts caused by implicit negation. Most existing approaches depend on explicit negation markers, leading to errors in context-dependent sentiment shifts. They lack domain adaptability, making them ineffective for specific applications like e-commerce reviews. Finally, they ignore syntactic variations, where negation is implied rather than directly stated (e.g., ‘The phone could have been better’ subtly implies dissatisfaction).

To address these shortcomings, this study develops a hybrid deep learning model integrating; Seq2Seq perturbations and contextual back translation for better data augmentation and context preservation. For the embedding purposes a hybrid of FastText and BERT embeddings is incorporated in order to enhance semantic

representation. Lastly, an attention mechanism is incorporated to improve polarity detection accuracy. The potential impact of this work is significant, not only for enhancing the accuracy of sentiment classification but also for improving real-world applications, such as consumer feedback analysis, market sentiment prediction, and recommendation systems.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter entails the methodology applied in this research study. A mixed research method was adopted to address the objectives of this research work. A systematic literature review research design was employed to conduct and analyse the literature review, focusing on identifying gaps and existing solutions used for handling the polarity shift due to implicit negation, in a qualitative manner. On the other hand, an empirical research design was used to develop and evaluate the proposed hybrid deep learning model in a quantitative manner. This study adopted the pragmatism research philosophy, combining both the qualitative and quantitative methods to achieve validation triangulation. In the research design, the overall strategy that guided the execution of this study was described. The experimental research process involved three key phases: the empirical baseline analysis, the model development and the model evaluation experiments. The model evaluation experiments aimed to evaluate the effectiveness of the newly developed hybrid deep learning model compared to the existing state-of-the-art sentiment analysis models used in the empirical baseline analysis process. The methodology followed a multi-methodological model development approach, starting with a conceptual design of the hybrid model, its architectural framework, and the implementation details. The rest of this chapter is organized as follows: research philosophy, research design, empirical baseline analysis, pilot study findings, experiments, integrated development environment, research instruments, sampling strategy, experimental data collection techniques, data analysis techniques, model development, and finally the research ethics.

3.2 Research Philosophy

This research study was based on the pragmatism research philosophy, employing a mixed research method for validation triangulation. The pragmatism philosophy emphasizes a choice of methods and designs that best address the research question, often by combining both qualitative and quantitative methodologies [134]. In this context, the study conducted experiments to evaluate the performance of existing sentiment analysis models in addressing polarity shifts. Experiments were designed

to systematically compare the existing models with the newly developed hybrid deep learning model. These evaluations relied on a carefully structured experimental checklist, which ensured consistency in the analysis of mobile phone brand reviews datasets. The use of both quantitative and qualitative metrics aligned with the pragmatism philosophy approach, as these metrics provide objective measures of model performance. By adhering to this philosophy, the study ensured that its findings were reliable, replicable, and grounded in measurable observations.

3.3 Research Design

This research adopted a mixed methods research design, integrating both qualitative and quantitative approaches as proposed by Creswell and Poth [135]. The qualitative component was used to explore the intricacies of handling polarity shifts in sentiment analysis caused by implicit negation in the written text. This involved a comprehensive analysis of the existing literature on sentiment analysis models, with a particular focus on approaches that handle the polarity shifts due implicit negation. During the literature review, the PRISMA approach proposed by Page et al. [136] was employed, providing a systematic framework for identifying, screening, and synthesizing relevant studies. The identification of the literature meta-search on articles focused on the “polarity shifters” and “sentiment analysis” as the main key words. A number of articles published between 2017 and 2024 were identified. Quality assessment and selection criteria further refined the choice of these articles to a lesser number. These articles were mainly from IEEE, ACM, research gate and springer databases. The final set of articles were reviewed, analysed and discussed. This process aimed at investigating the critical polarity shifters in sentiment analysis within written text. A pilot-study of the best performing sentiment analysis models for handling polarity shift was used to conduct quantitative research in the second step. The pilot study was utilized to evaluate the performance of the sentiment analysis model that best handles polarity shifts from the theoretical analysis. This validated or challenged the claims contained in the theoretical analysis. The quantitative component focused on systematically collecting and analysing numerical data to evaluate the performance among the existing models as well as against the novel hybrid model. This involved designing and conducting experiments to test hypotheses, answer research questions, and draw statistical inferences.

Experimental data was collected following a structured experimental methodology, ensuring consistency and reliability in the results**.** To achieve the third and fourth objective, this study utilized an experimental research approach that integrated both qualitative and quantitative methods to create and validate a novel hybrid deep learning model. The results from the experiments were tabulated and statistically analysed using Jamovi, an interface of the R software. Jamovi was chosen for its intuitive point-and-click interface, which simplified performing advanced statistical analyses such as ANOVA. This software was particularly advantageous for this research, as it reduced the time required for complex statistical analyses compared to Python, which often involved extensive coding [137]. The statistical analysis outcomes guide the refinement and validation of the hybrid model.

In the validation process, a combination of both the quantitative and qualitative methods was employed. Statistical techniques such as ANOVA and the Kruskal Wallis H-statistic were used to quantitatively evaluate the model's performance. On the qualitative side, visualizations of the attention heatmaps and whisker plots provided insights into the model's ability to handle polarity shifts, due to implicit negations, in a variety of mobile phone reviews datasets. The SHAP visualization was also used in order to provide granular insights into the input feature contribution to the prediction outcome. This multi-method validation approach ensured that the research was both rigorous and comprehensive. The alignment between qualitative and quantitative methods was assessed to ensure coherence and complementarity in addressing the research objectives. Practical constraints such as computational resources and ethical considerations in the use of publicly available sentiment data were carefully considered. These measures ensured the research process is efficient, focused, and capable of generating valuable insights into addressing polarity shifts due to implicit negations, in sentiment analysis.

The mapping of research objectives to the respective methods ensured a structured approach to addressing the research problem. This mapping process involved identifying appropriate methods to evaluate the model's performance, optimize its design, and validate its effectiveness in real-world sentiment analysis tasks. Figure 3.1 illustrates the mapping of research questions to their respective methods, as guided by the framework proposed by Järvinen [138].

This alignment underscored the systematic nature of the research design in achieving the stated four research questions.

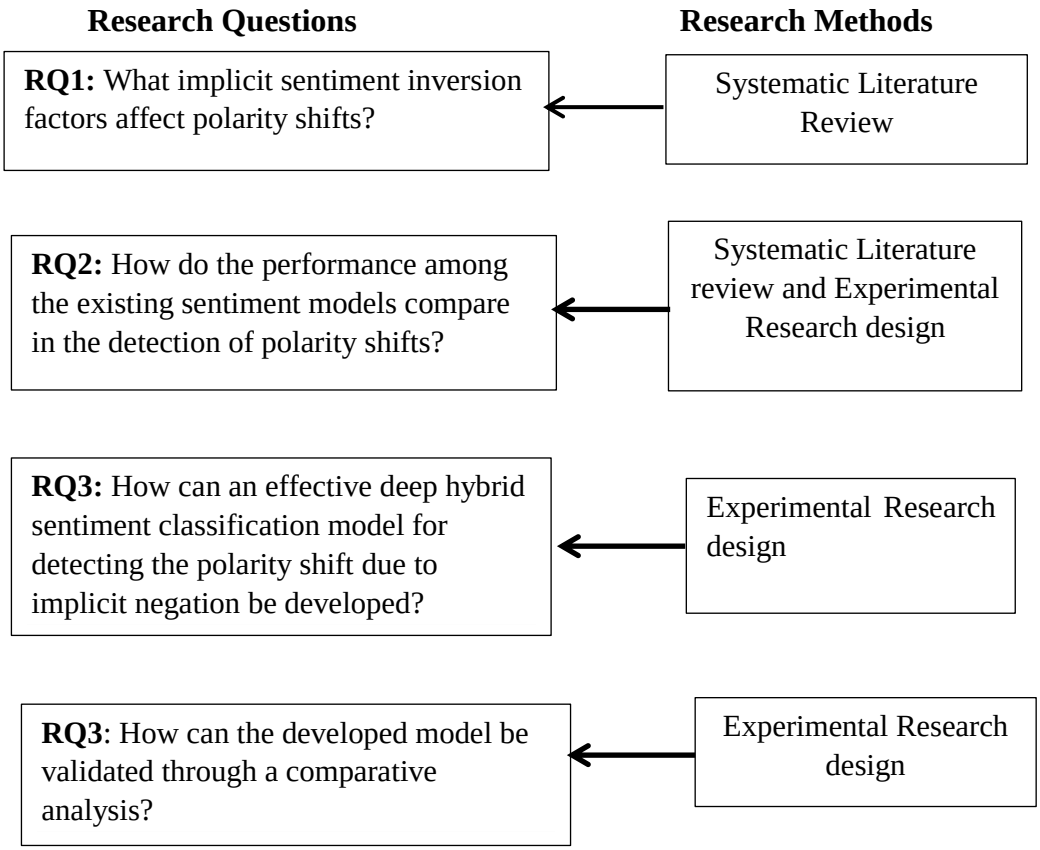


Figure 3.1: Mapping of research objectives, and questions to research methods

3.3.1 Systematic Literature Review Methodology

The systematic literature review in this study involved a systematic exploration and theoretical analysis of existing scholarly works, focusing on sentiment analysis models addressing polarity shifts due to implicit negation. The theoretical analysis is pivotal in identifying gaps in current approaches and informing the framework for developing the hybrid deep learning model proposed in this study. A structured meta-search was conducted across databases such as IEEE, ACM Digital Library, SpringerLink, and ResearchGate. The review targeted peer-reviewed publications from 2017 to 2024 with keywords like "implicit negation," "polarity shift," and "sentiment analysis models." This review followed PRISMA approach of the systematic literature review methodology including identification, inclusion and screening of the relevant studies. This iterative approach ensured a robust alignment between insights of the theoretical analysis and experimental validation.

3.3.2 Pilot Study

The pilot study served as a preliminary step to refine the experimental checklists for the empirical baseline analysis process adopted for evaluating sentiment analysis models' ability to handle polarity shifts due to implicit negation. This study aimed to assess the feasibility of the experimental design, validate the selection of datasets, and test the performance of a small subset of existing sentiment analysis models under controlled conditions. Models identified during the systematic literature review phase were implemented using Python in a cloud-based environment, such as Google Colab, to simulate real-world scenarios involving mobile phone brand reviews. Accuracy was applied as the key performance metric in evaluating the effectiveness of these sentiment analysis models in handling polarity shifts due to implicit negation. The insights from the pilot study informed adjustments to the experimental and observations checklists, for example dataset preprocessing strategies and parameter configurations for model comparisons**.** The best performing sentiment analysis models that exhibited consistency in both the theoretical analysis and the empirical performance in the pilot study formed the sample for the empirical baseline analysis in the main experiments.

3.3.3 Experimental Design for the Empirical Baseline Analysis

The experimental research design in this study focused on planning and conducting experiments aimed at investigating the end-to-end behaviour, and the interplay of sentiment analysis models used for handling polarity shifts due to implicit negation. In this empirical baseline analysis, the existing best performing techniques used in the main experiments were identified through both the theoretical analysis process and the pilot study. By subjecting these models to a benchmark mobile phone reviews dataset containing implicit negations, and in a variety of feature extraction, feature selection and data augmentation techniques, the objective of the experiments in this research study aimed at scrutinizing their efficacy in accurately capturing sentiment shifts caused by implicit negation. This helped uncover the differences in their performance in adapting to the dynamic nature of sentiment expression and polarity shift due to implicit negation. This exploration paved way for advancements in sentiment analysis methodologies, fostering the development of improved a novel model that could adeptly navigate the intricate landscape of sentiment flux induced by implicit negation. Based on the experimental design, the research hypothesis was

first formulated in line with the research questions, before conducting the pilot study and the experiments. This empirical baseline analysis examined the performance among the existing best-performing sentiment analysis models. The identification of the existing models used in the experiments contained in the empirical baseline analysis was based on insights from a thorough theoretical analysis and the pilot study.

In the subsequent sections, the Integrated Development Environment (IDE) for conducting the experiments is described. The sampling strategy used in the selection of the input mobile phone reviews datasets, sentiment analysis models, feature extraction techniques, feature selection techniques and data augmentation techniques used in the various experiments is also described. Lastly, the experimental data collection process and techniques are discussed, including the data analysis techniques for the experimental data.

3.3.3.1 Main Experiments

The main experiments focused on rigorously evaluating sentiment analysis models to address the challenges posed by polarity shifts due to implicit negation. Five categories of models were tested, i.e., traditional machine learning (ML) Models, deep learning models, hybrid models combining traditional ML and deep learning, hybrid Models using exclusively deep learning approaches, and hybrid models using exclusively traditional ML approaches. The specific deep learning models included BERT (Bidirectional Encoder Representations from Transformers) and Long Short-Term Memory (LSTM) while the specific traditional machine learning models included Naïve Bayes, and Support Vector Machines (SVM). For feature extraction, state-of-the-art methods such as FastText and Word2Vec were employed, leveraging their robust ability to capture contextual and semantic relationships. Recursive Feature Elimination (RFE), a wrapper-based technique, was used for feature selection, proving more effective than traditional methods like TF-IDF in optimizing the feature space. Data augmentation techniques included back translation and synonym replacement, which enhanced the dataset by introducing semantic variability without altering the core sentiment of the text. The choice of these experimental variables was in line with the observations made in both the systematic review paper by Murithi et al. [124] as well as in the empirical analysis paper by

Murithi et al.[139]. The main experiments utilized the Amazon Mobile Phones Reviews dataset, selected for its rich representation of sentiment variability, including instances of polarity shifts due to implicit negations and contextual cues. This dataset, available from the Kaggle repository, offered diversity in language styles, user demographics, review lengths, and polarity patterns, making it ideal for testing advanced sentiment analysis models. During preprocessing, the feature extraction, feature selection, and data augmentation techniques were systematically varied in each experiment to explore their combined effects on model performance. By adopting this structured approach, the research provided a robust evaluation of the strengths and limitations of current sentiment analysis methods, ultimately guiding the development of the proposed hybrid deep learning model tailored to effectively handle implicit negations.

The proposed model architecture was designed into five layers that were defined as Input layer that accepts the concatenated embeddings from the feature extraction defined as matrix F_refined. Attention mechanism layer applied a Scaled Dot-Product Attention to highlight the most critical sentiment tokens in the data. Feed forward layers included two dense layers, each with 128 neurons and ReLU activation while the pooling layer performed mean pooling over token embeddings for sentence representation**.** Lastly the output layer contained a fully connected softmax layer for multi-class sentiment classification. The model training process involved a number of steps. Firstly, training dataset labelled as D_train and hyperparameters H that included: 32 Batch size, 2e-5 learning rate with a dynamic scheduling, 0.3 dropout rate, Cross-Entropy loss function and Adams optimizer acted as the main input in the training process. The model is trained for 40 epochs while monitoring the validation loss. Early stopping was set in cases where there were no improvements in 3 consecutive epochs. The trained model was tested using validation metrics that included: accuracy, Cohen’s kappa, Mean Squared error, and Matthews Correlation Coefficient (MCC). Error cases are analysed with the main focus on misclassified implicit negations. This ensured the model’s sentiment classifications were accurate and could be trusted by business to make decisions.

3.3.3.2 Population Sampling Strategy and Sample Size

In line with the purposive sampling strategy proposed by Etikan et al. [140], this research adopted a systematic approach to selecting variables for use in the main experiments of the empirical baseline analysis process. Based on the outcomes of the theoretical analysis and the pilot study, the study focused on a subset of sentiment analysis models, feature extraction methods, feature selection techniques, and data augmentation methods that demonstrated strong potential for handling polarity shifts caused by implicit negation. From an initial population of eighteen sentiment analysis models, four are selected, including BERT, LSTM, Naïve Bayes, and SVM. For feature extraction, two methods are selected from a pool of six i.e. FastText and Word2Vec were chosen from a pool of nine. The choice of the feature selection techniques narrowed down to both the Recursive Feature Elimination (RFE) and the TF-IDF (Term Frequency-Inverse Document Frequency) from a pool of four potential techniques.

Lastly, two data augmentation techniques, including back translation and synonym replacement, were selected from five available techniques. The use of purposive sampling is cost-effective and time-efficient compared to random sampling strategies, aligning with the propositions of Banerjee et al. [140] . The research incorporated benchmark datasets, particularly the Amazon Mobile Phones Reviews dataset due to its robustness in capturing real-world sentiment variability, including implicit negation. The focus on publicly available benchmark datasets enhanced reproducibility and reliability, providing a common platform for evaluating the effectiveness of the selected models, features, and augmentation methods. This focused sampling approach was driven by the need for quality as opposed to quantity, as the study targets methods and techniques with proven methodological and evaluation rigor, high level of novelty and significance, as well as the practical relevance. This helped achieve a balance with satisfactory analytical depth and statistical power that befitted this kind of multifaceted exploration in this kind of research study. By concentrating on a manageable subset of variables, the research achieved resource efficiency and computational feasibility while ensuring statistically significant outcomes. The purposeful selection of variables and datasets enabled deeper analysis, paving the way for meaningful contributions to sentiment

analysis, particularly in addressing the challenges of polarity shift due to implicit negation.

i. Sample Models

Four sample sentiment analysis models used for handling polarity shifts due to implicit negation were selected for the main experiments of the empirical baseline analysis. These include: SVM for model 1, BERT and NB for model 2, BERT and LSTM for model 3, and SVM and NB for model 4. This selection was guided by their state-of-the-art performance and unique capabilities in handling polarity shifts caused by implicit negation evinced in the theoretical analysis and the results of the pilot study as well as in the works of Murithi et al [124] and Murithi et al [139]. Besides the works of Murithi et al [124]and Murithi et al [139], it is also evident that the Support Vector Machines (SVMs) have long been established as a robust traditional machine learning model for text classification, particularly in sentiment analysis. The strength of SVM lies in its ability to find the optimal hyperplane that maximizes the margin between classes, making it highly effective for binary and multi-class problems [141]. Additionally, its generalization capabilities allow it to effectively capture implicit negation cues when combined with well-engineered features [142].

The hybrid model combining BERT (Bidirectional Encoder Representations from Transformers) with Naïve Bayes (NB) leveraged the contextual embeddings of BERT with the probabilistic simplicity of NB. BERT, developed by Devlin et al.[143], is a transformer-based language model that excels at understanding contextual relationships in text. Its bidirectional training process allows it to capture subtle polarity shifts caused by implicit negations, which often depend on the surrounding textual context. The probabilistic nature of NB complements BERT's embeddings, enabling it to capture nuanced sentiment shifts, including those driven by negation markers [144].

This is also evinced in the hybridization of BERT and LSTM represents a synergy between BERT's contextual representation capabilities and Long Short-Term Memory (LSTM)'s ability to model sequential dependencies.

LSTM, introduced by Hochreiter and Schmidhuber [145], is adept at handling sequential data, which is critical for understanding polarity shifts arising from implicit negation embedded in textual sequences. When integrated with BERT's pre-trained embeddings, the model achieved a superior ability to capture both context and temporal relationships. This combination is particularly suited for long-text reviews, where polarity shifts might span across multiple sentences or phrases [146].

Lastly, it is also evinced that the hybridization of SVM and Naïve Bayes merged the strengths of both models, leveraging SVM's discriminative power with NB's generative probabilistic reasoning. This approach has been shown to improve classification performance in sentiment analysis tasks, especially when dealing with noisy or imbalanced datasets [147].

ii. Datasets

The datasets selected for the experiments included the Amazon Mobile Phone Reviews, Google Play Store Reviews, and iPhone Apple App Store Reviews. These datasets offered rich sources of diverse user-generated content that enabled a comprehensive evaluation of sentiment analysis models for handling polarity shift due to implicit negation. Each of these datasets brought unique aspects that contributed to the generalizability and robustness of the models. Amazon Mobile Phone Reviews is widely regarded as a benchmark in sentiment analysis research due to its large scale, diversity in language, and the presence of complex sentiments that may shift due to implicit negation [148]. Amazon mobile phone datasets contained 67,987 records and are annotated with different scores value. Negative sentiment polarity is annotated with values ranging between 1 and 2. The positive sentiment polarity is annotated with values ranging between 4 and 5. Lastly the neutral sentiment polarity is annotated with value 3. The dataset contained a number of features including Serial number to label the record number, Name of the mobile phone, Rating, Date, title, body and votes. This dataset contains detailed product reviews that included both positive and negative sentiments, often with nuanced linguistic features like negation, making it ideal for studying polarity shifts [139].

Both Google Play Store Reviews and iPhone Apple App Store Reviews used during the hybrid model evaluation were annotated with different values for negative, neutral and positive polarities. Google play store reviews contained all the details of the applications on Google Play. It had thirteen features that describe a given app and contained contains 100 reviews for each app. The text in each review was pre-processed and annotated with three sentiments (Positive, Negative or Neutral).iPhone Apple App store reviews contained reviews about the Apps on iPhone Apple Play Store. The reviews provided information about the Apps like App_Name, Primary_Genre in which they are launched and their target audience age groups. Moreover, the user reviews provided information about the developers of the Apps, their reviews, average user rating, and size of the apps in bytes and price of the Apps. This data frame consisted of 1230375 rows. The annotation pairwise agreement was measured using Cohen's kappa. This ensured the validity of the datasets was used in this study.

Additionally, its availability and extensive metadata, such as product attributes and user demographics, further enriched its utility for model training. The Google Play Store Reviews dataset offered another diverse set of user-generated content, focusing on app feedback, where users often provide feedback on their experiences with implicit sentiments [149]. This dataset's value lies in its varied vocabulary, different app categories, and the presence of polarity shifts in reviews where users comment on their experiences negatively but under certain conditions, like app updates or bugs, which could introduce negation. These features enable the exploration of implicit negation and the associated polarity shifts in sentiment analysis. Lastly, the iPhone Apple App Store Reviews dataset provides additional complexity, as it includes diverse user opinions on iOS apps, where indirect language and implied sentiments are common [150]. Its richness in implicit negation allowed for testing models in a setting with frequent sentiment shifts caused by issues such as updates, bugs, or performance degradation. The dataset's structured metadata, such as user ratings and app categories, further guided model evaluation and ensured that polarity shifts were well-represented.

The generalization of the novel model was strengthened by utilizing multiple datasets in the validation process. This generalization ensured that the sentiment analysis model could handle a broad range of real-world scenarios where implicit negation might cause shifts in polarity. This diversified approach to dataset selection and validation ensured that the findings were not overly specific to a single dataset, which was essential for validating models and to confirm that they perform well across various datasets. By leveraging datasets with varied characteristics, this research ensured that the proposed model could robustly handle implicit negation and polarity shifts, making it highly adaptable and applicable to real-world sentiment analysis challenges in relation to mobile phone reviews and beyond.

iii. Feature Selection Methods

In this research study, two widely used feature selection techniques i.e. Recursive Feature Selection (RFE) and Term Frequency-Inverse Document Frequency (TF-IDF) were used in line with the empirical works of Murithi et al [139]. These techniques were particularly effective due to their ability to prioritize relevant features and reduce dimensionality. In sentiment analysis models, particularly when addressing polarity shifts due to implicit negation, the choice of feature selection techniques was crucial for improving model performance and interpretability. RFE operates by training the model with all features, evaluating the importance of each feature, and eliminating the least significant features one by one until the optimal set is identified [151]. RFE has been found effective in sentiment analysis tasks, especially when dealing with implicit negation, because it allows the model to focus on the most relevant features (such as key words or phrases) that affect sentiment polarity, including negations. By removing redundant or irrelevant features, RFE helps improve the model's efficiency and generalization, thus enhancing its ability to capture shifts in polarity caused by implicit negation in text [152].

On the other hand, Term Frequency-Inverse Document Frequency (TF-IDF) is a classic text feature extraction method that weighs the importance of each word in a document relative to its frequency across a corpus [153]. It has been widely used in sentiment analysis, particularly for capturing important

keywords that affect sentiment. TF-IDF is advantageous in handling polarity shifts due to implicit negation because it gives higher weight to terms that are more unique to the document while downplaying common terms across all documents. The ability of TF-IDF to emphasize terms that are statistically significant within a specific context makes it an essential feature selection tool for analysing sentiment changes induced by implicit cues [154].

The combination of RFE and TF-IDF ensured that the model selected only the most important features that contribute to sentiment shifts. While RFE helped eliminate irrelevant features and prevent overfitting, TF-IDF allowed the model to prioritize the most relevant terms for each specific sentiment analysis task, including those that reflect polarity shifts due to negation.

iv. Feature Extraction

Based on the empirical works of Murithi et al [139], both the FastText and Word2Vec were the two feature extraction techniques used in this research study. The two are also the widely adopted word embedding techniques as they effectively capture relationships, making them suitable for handling sentiment shifts, including those caused by implicit negation. In sentiment analysis models, particularly those addressing polarity shifts due to implicit negation, feature extraction played a vital role in capturing the semantic relationships between words and their context. FastText, developed by Facebook AI Research, improves upon traditional word embedding models like Word2Vec by representing each word as a bag of character n-grams [155]. This allows FastText to capture subword information, making it particularly useful for handling rare words or words with multiple meanings, such as negations (e.g., "not good" vs. "good"). FastText's ability to model out-of-vocabulary (OOV) words by using character-level representations helped in handling the complex semantic shifts caused by implicit negation in text. Since implicit negation often involves subtle shifts in sentiment that may depend on the presence of specific negation words or affixes, FastText can represent these variations effectively, providing more accurate sentiment classifications [156]. Its robustness to morphological variations and its ability to handle rare words

made FastText an ideal choice for sentiment analysis, especially in the context of analysing polarity shifts triggered by negation cues.

Word2Vec, developed by Mikolov et al.[157], is another widely used word embedding technique that represents words in a continuous vector space.Word2Vec works by mapping words into a dense vector space such that words with similar meanings are closer to each other. This is accomplished through either the Continuous Bag of Words (CBOW) or Skip-gram model, which predict a word based on its context or vice versa. Word2Vec has proven highly effective in capturing semantic meaning and context, which is crucial when analysing how negation words (e.g., "not," "never," "no") can shift the polarity of a sentence. By capturing the relationships between words in context, Word2Vec helped in distinguishing between positive and negative sentiment in the presence of negation, which is often subtle and implicit. Its ability to map words into a semantic space where similar words are closer together makes it effective for capturing the nuances of sentiment and detecting polarity shifts, which are common challenges in sentiment analysis[157]. Both FastText and Word2Vec were suitable for addressing the challenges posed by implicit negation in sentiment analysis due to their ability to capture contextual meaning and semantic shifts. The choice of these techniques is grounded in their ability to represent words and their contexts effectively, ensuring that the sentiment shifts caused by negation can be detected and understood.

v. Data Augmentation Techniques

Based on the empirical works of Murithi et al [139], both the back translation and the synonym replacement were the two data augmentation techniques used in this research study. In deep sentiment analysis models, especially those designed to handle polarity shifts due to implicit negation, Back Translation and Synonym Replacement were two data augmentation techniques that significantly improved the model's ability to generalize and handle complex shifts in sentiment. These methods helped the model to better understand and adapt to variations in sentiment caused by negation and contextual cues.

Back Translation is a technique where a sentence is translated into another language and then translated back into the original language. This technique has gained prominence for its ability to create diverse linguistic variations of a given sentence while maintaining its semantic meaning [158]. Back translation can be particularly effective in enhancing sentiment analysis models by generating paraphrased sentences that may include implicit negation or shift in sentiment, thus allowing the model to better handle polarity changes. For example, the phrase "I do not like this product" could be back-translated into "I dislike this product," providing the model with different sentence structures that express the same sentiment. This augmentation is crucial in models dealing with implicit negation, as it exposes the model to multiple ways in which negations can be expressed, improving its ability to detect shifts in sentiment due to negation cues.

Synonym Replacement is another widely used technique in data augmentation where specific words in a sentence are replaced with their synonyms. This method helps the model recognize the same sentiment despite variations in vocabulary, which is particularly useful in handling implicit negations[159]. For instance, replacing "good" with "excellent" or "bad" with "terrible" helps the model to learn that despite these variations in wording, the overall sentiment remains the same, or it can help to identify subtle shifts in polarity when combined with negation words. By incorporating synonyms, the model becomes more robust to changes in language usage, which is essential for understanding how negation or other contextual clues can affect sentiment. This technique has been shown to improve the performance of deep learning models by enriching the training data and helping the model generalize better to unseen text, especially when dealing with more complex and varied expressions of sentiment[160].

Both Back Translation and Synonym Replacement are powerful techniques for augmenting sentiment analysis datasets. They provide the model with diverse examples of how sentiment can be expressed, enhancing the model's ability to handle nuanced changes in sentiment caused by implicit negation. These augmentation methods not only help in generating diverse variations of the

same sentiment but also improve the model's robustness to different linguistic structures, making them highly effective in improving sentiment classification accuracy. Figure 3.2 is an activity diagram that represents the empirical analysis on the sentiment analysis model in a variety of feature selection, feature extraction and data augmentation techniques.

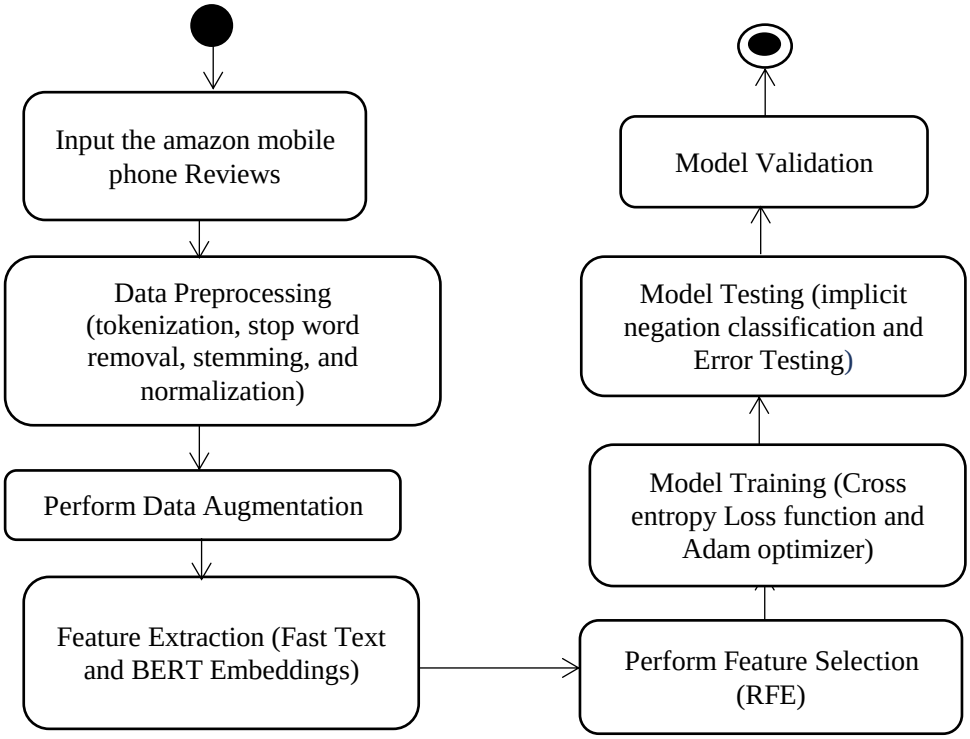


Figure 3.2: Activity Diagram for the Empirical Analysis

The experiments of this research involved a number of processes. Firstly on amazon mobile phone reviews datasets, a number of preprocessing so as to remove any noise. Data augmentation using back translation and synonym replacement was then performed on the datasets in order to improve the data for the training process. Vectorization was done using both the FastText and word2vec to prepare data for the training process. Feature selection through the use of wrapper based recursive feature elimination and Term Frequency Inverse Document Frequency (TF-IDF) techniques followed. In each experiment, accuracy, Mathews' coefficient correction and Cohen's kappa were recorded as the performance metrics. Detailed overview of the unique experimental combinations designed for the empirical analysis of baseline models, incorporating various categories of sentiment analysis models, feature extraction techniques, feature selection methods, and data augmentation techniques

is shown in appendix 2. These experiments aimed to systematically evaluate the performance of different combinations to identify the most effective configuration for sentiment polarity classification.

The sentiment analysis models under consideration included individual traditional machine learning models i.e. SVM and Naïve Bayes, individual deep learning models i.e. BERT and LSTM, hybrid models combining traditional models only i.e. SVM + Naïve Bayes, hybrid combining deep learning models only i.e. BERT + LSTM, and hybrid models that integrated both traditional machine learning and deep learning models i.e. SVM + BERT, SVM + LSTM, Naïve Bayes + BERT, Naïve Bayes + LSTM. For feature extraction, the study utilized FastText and Word2Vec, while feature selection techniques included Recursive Feature Elimination (RFE) and TF-IDF. Additionally, the study employed data augmentation techniques such as back translation and synonym replacement.

3.3.3.3 Research Instruments

In this study, several research instruments were employed to ensure a rigorous and systematic approach to both the experiments and the data analysis. One of the primary tools used is the experimental checklist shown in appendix 1, which played a pivotal role in organizing and documenting the various experimental steps and configurations used throughout the research process. This checklist ensured that each experimental setup, ranging from model selection to feature extraction techniques, were consistently followed, allowing for clear tracking of methodologies and enabling reproducibility [161]. Coupled with the experimental checklist, the observation checklists were utilized to monitor and document key aspects during the experimentation phase. The observation checklist provided a detailed guide for tracking the performance of the models, using the predefined evaluation metrics. This tool is crucial for maintaining objectivity and consistency when analysing experimental outcomes, ensuring that all observations are systematically recorded and reviewed [162]. Another essential instrument used is data preprocessing scripts, which are vital for preparing the raw data for analysis. Data preprocessing, including tasks such as normalization, handling missing values, and feature encoding, is crucial in machine learning to improve model accuracy and reliability [163]. These scripts automated the preprocessing workflow, ensuring that the same procedures were

applied consistently across the various experiments used in this study. This approach minimized human error and ensured the reproducibility of results by maintaining uniform preprocessing steps.

The data analysis software and statistical tools were employed as the research instruments for conducting robust analyses of the experimental results. The R-based data analysis software enabled the computation of the ANOVA in order to help evaluate the significance of model performance across different experimental setups [146]. Python software together with the respective libraries were utilized for implementing machine learning models. These research instruments allowed for the efficient execution of algorithms, statistical tests, and performance evaluations, producing reliable and reproducible results. By integrating these research instruments, this study was able to maintain a high level of methodological rigor, transparency, and consistency, ensuring that the research findings were both valid and reproducible.

3.3.3.4 Integrated Development Environment and Libraries

The integrated development environment (IDE) for this research study mainly included Jupyter Notebook on Google. This IDE offered seamless integration with a number of essential libraries, enabling efficient coding, testing, and evaluation of models. Table 3.1 describes the summary of all the libraries used in this research work.

Table 3.1: Libraries in Different Model Categories

Library/Tool	**Purpose**
Scikit-learn	Traditional ML models (SVM, Naïve Bayes)
Transformers	Pre-trained BERT model for deep learning-based sentiment analysis
Torch	Deep learning framework for training and fine-tuning LSTM and BERT models
Scikit-learn, Transformers, Torch	Combined traditional ML (SVM, Naïve Bayes) with deep learning models (BERT, LSTM)
Fasttext	FastText for word embeddings
Gensim	Word2Vec for word embeddings
Scikit-learn	Recursive Feature Elimination (RFE), TF-IDF
Googletrans	Back translation for text augmentation
Nltk, Wordnet	Synonym replacement to augment dataset
TensorFlow, Keras	Model training and deployment for deep learning-based tasks
PyTorch	Flexible deep learning framework for complex architectures (BERT, LSTM)

Jupyter Notebook is particularly popular due to its interactive nature, allowing for easy visualization and step-by-step experimentation with models and datasets. SVM required two key libraries i.e. Scikit - learn and Numpy. Scikit-learn are widely used for implementing Support Vector Machines (SVMs) due to its efficient algorithms for classification and its comprehensive suite of machine learning tools. It provides robust implementations of SVM, enabling model training and evaluation with minimal computational overhead [164]. Numpy is used for matrix operations and data manipulation. Support Vector Machines (SVM) are a well-established method for text classification and sentiment analysis, especially for tasks involving implicit negation [141]. Scikit-learn offers an easy-to-use implementation of SVM with hyperparameter tuning, enabling fine-tuning for different datasets, which is critical for handling the sparse features and non-linear nature of sentiment-rich text [142].

The hybrid model of BERT and Naïve Bayes Model required transformers from hugging face, torch and scikit-learn. Transformers (from Hugging Face) provides a pre-trained implementation of BERT [143], making it easy to use BERT embeddings for textual sentiment analysis. The transformer was well-suited for capturing

complex contextual relationships and handling polarity shifts architecture due to implicit negation. Naïve Bayes (NB) was implemented in Scikit-learn, which provided an efficient, simple probabilistic model for sentiment classification. Torch was necessary for BERT model training and optimization when leveraging BERT's pre-trained embeddings. BERT's bidirectional architecture enables the model to capture subtle polarity shifts arising from implicit negation, while NB enhances the ability to handle probabilistic relationships in textual data [144].

This hybrid approach provided a balance between understanding nuanced contextual cues and maintaining computational simplicity. The hybrid of BERT and LSTM required transformers from hugging face, Keras and torch. Transformers (from Hugging Face) were used for BERT’s contextual embeddings, and were essential for fine-tuning BERT models and leveraging pre-trained weights. Keras (or TensorFlow) used with LSTM are a type of recurrent neural network (RNN) library, which provided tools for creating and training LSTM layers in a deep learning model. Torch was used for integrating BERT's embeddings with LSTM in a PyTorch environment, ensuring compatibility with GPU acceleration for faster training.

For feature extraction, the study utilized FastText, a library developed by Facebook for efficient word embeddings, and gensim, which provides an implementation of Word2Vec, another popular method for generating word embeddings that capture semantic relationships between words. Feature selection was handled through scikit-learn, which offers techniques like Recursive Feature Elimination (RFE) and TF-IDF, allowing the study to focus on the most relevant features in the dataset.

Data augmentation was performed using the googletrans library for back translation, which translates text into another language and then back into the original language to introduce variability, and nltk along with wordnet to implement synonym replacement, a technique that substitutes words with their synonyms to further augment the dataset. In addition to these libraries, deep learning and model deployment were supported by frameworks such as TensorFlow and PyTorch, which provided the necessary infrastructure for training and deploying models at scale. TensorFlow, used in conjunction with Keras, enabled efficient model training for deep learning tasks, while PyTorch, known for its flexibility, supported the dynamic computational graphs required for complex architectures like BERT and LSTM.

Together, these libraries formed the backbone of the study, offering optimized tools for building, training, and evaluating the different sentiment analysis models,

The hybrid model of BERT and LSTM model combined the power of BERT's context understanding with LSTM's ability to model sequential dependencies. This approach improved performance on tasks that involved complex sentence structures and multi-sentence polarity shifts. The complementary strengths of SVM and NB ensure high accuracy when detecting nuanced sentiment shifts, especially in datasets with noisy or imbalanced samples [147];[16].

The hybrid novel model combining BERT and LSTM with Contextual Back Translation Augmented by Seq2seq Perturbations required a number of libraries. Transformers (from Hugging Face) were used for leveraging pre-trained BERT models, which provided contextual embeddings that captured the nuanced meaning of words in context. Keras (or TensorFlow) were used for training the LSTM layer and combining it with BERT's embeddings. Fairseq (or OpenNMT) was used for implementing sequence-to-sequence models and handling back-translation and perturbation tasks. The Google Translate API was used as the Back-Translation API in order to augment training data via back-translation techniques. The hybrid model of BERT and LSTM model, augmented by contextual back translation and Seq2seq perturbations, is an advanced method for handling polarity shifts due to implicit negation. By leveraging back-translation, this model enriched the training dataset with multiple variations of the same sentiment, capturing different linguistic structures and nuances of negation [158]. The Seq2seq perturbations further enhance the model's robustness by introducing additional variations of the training data [139].

3.4 Experimental Data Collection and Techniques

The experimental data collection process integrated the experimental checklists to ensure a systematic approach to deriving performance metrics from various sentiment analysis models executed on Amazon Mobile Phone review dataset. This process began with preprocessing the dataset to ensure consistency, applying techniques such as tokenization, stemming, lemmatization, and handling special characters[166]. In the experiments tokenization was done using BERT's WordPiece tokenizer that ensured contextual embedding compatibility. FastText embedding was

done for standard tokenization while normalization of text was done through lower casing and lemmatizing words. Additional data preprocessing was done through augmentation in order to improve the data for training purpose. In this research work, tokenization and normalization handled variations in linguistic patterns, crucial for accurately capturing the context of negations. This was done using back translation augmented by Seq2Seq perturbation through a Seq2Seq encoder i.e. MarianMT and decoded to multiple syntactic variations using beam search. Lastly validation was done on the preprocessed data to ensure semantic consistency, retention of domain specific terms and idiomatic expressions. Feature extraction methods were utilized to transform the textual data into numerical representations, while feature selection techniques optimized the feature space [157]; [167];[143]. Data augmentation techniques were applied to mitigate the effects of polarity shifts caused by implicit negation [158].

Performance metrics such as accuracy, Cohen's kappa, Matthews's correlation coefficient (MCC), and mean squared error (MSE) were then recorded following the execution of these models. Experimental checklists ensured all required configurations such as hyperparameter tuning, feature engineering, and augmentation were consistently applied across all experiments. Observation checklists document any anomalies, computational challenges, or noteworthy behaviours during model training and evaluation, ensuring comprehensive monitoring of the experimental process. To ensure reliability, experiments were repeated under varying train-test splits. K-fold cross-validation approach further evaluated model performance across multiple data partitions [168].

Statistical reliability was assessed using Cronbach's alpha, a measure of internal consistency, where values exceeding 0.7 indicate strong reliability in the collected data [169]. This ensured consistency and repeatability of results across different configurations and experiments. Validity testing was an essential component of the data collection process. Construct validity was ensured by designing experiments aligned with established principles in sentiment analysis research, confirming that the metrics reflected the models' ability to handle polarity shifts. Content validity was achieved through the use of well-recognized datasets and benchmarking techniques, enabling comparisons with state-of-the-art models. Criterion-related

validity was tested by comparing baseline models with the hybrid model, validating the improvements achieved[16]. Face validity was further reinforced through observation checklist confirming the appropriateness of the methodology for addressing the research objectives. Together, these measures ensured that the data collection process was systematic, reliable, and valid, providing robust evidence to evaluate the effectiveness of the proposed hybrid model.

3.5 Experimental Data Analysis

The experimental data analysis for this research study was designed to comprehensively evaluate the performance of sentiment analysis models on mobile phone review datasets. The analysis incorporated both quantitative and qualitative approaches to assess the performance of the existing state of the art sentiment analysis models among themselves and against the hybrid deep learning model, in their ability to handle polarity shifts caused by implicit negation. Quantitative analysis began with the computation of performance metrics, including accuracy, Cohen's kappa, Matthew's correlation coefficient (MCC), and mean squared error (MSE). These metrics provided insights into various dimensions of model performance, such as correctness, agreement beyond chance, and error magnitude. These metrics were combined into a single universal metric as shown in the following equation:

$$\text{average normalized perfomance score} =$$
$$(\text{normalized accuracy} + \text{normalized cohen}'\text{s kappa} +$$
$$\text{nomarized Mathews correlation})/3)$$

Combining these metrics into a single metric was grounded in the principle of creating a balanced and holistic evaluation metric that captures diverse aspects of model performance, with each addressing specific facets of performance, and their combination ensured a comprehensive assessment [144]. On the other hand, the normalization of individual performance scores was essential for ensuring comparability across the various evaluation metrics, and ensured that they operated on similar numerical scales, which can led to imbalances when comparing their raw values.

Normalization scaled these scores to a common range, typically 0 to 1, allowing them to be combined and compared without one metric dominating due to scale differences. This process eliminated biases introduced by raw scale variations and ensured that the evaluation provides a balanced view of model performance [170]. The formula for normalization is shown below:

$$x_{norm} = \frac{\boldsymbol{x - min(x)}}{\boldsymbol{max(x) - min(x)}}$$

The equation above transforms the original value x into a new value x_norm that falls within the range [0, 1].

The above symbols are explained as follows:

x = the original value you want to normalize.

min(x) = The smallest value in the dataset or feature column.

max(x) = The largest value in the dataset or feature column..

During the validation of the novel model, statistical methods like the Kruskal-Wallis H test were applied to determine the significance of the difference of the novel hybrid model against the baseline models. Cross-validation techniques, such as k-fold cross-validation, further ensured that the empirical analysis results were generalizable across diverse dataset splits, minimizing the risk of overfitting or data dependency. Qualitative analysis focused on the interpretability of the hybrid deep learning model compared to the existing baseline models. Attention heatmaps were employed to visualize the model's focus on critical words or phrases that influenced sentiment classification, particularly in sentences affected by implicit negation. Whisker plots were utilized to graphically represent the distribution and variability of performance metrics across various configurations, offering a comparative visual summary of the models**.** Additionally, observation checklists were used to capture anomalies, biases, or patterns in model predictions, providing deeper insights into areas requiring refinement in these models. The insights gained from the experimental data analyses were integrated into the development of the hybrid model. For instance, the two way ANOVA statistical results informed hyperparameter configurations based on the interactions.

Lastly, the validation exercise of the novel hybrid model adopted a holistic evaluation of the hybrid model's capability to address polarity shifts due to implicit negation using both the quantitative and the qualitative strategies.

3.6 Research Ethics

Research ethics are fundamental principles that guide scientific inquiry, ensuring integrity, responsibility, and respect for data and the research process [171]. Ethical considerations are vital at all stages of research, from the design phase through data collection, analysis, interpretation, and reporting, in order to maintain the integrity of the findings and uphold public trust in the research community [172]. In this study, honesty and transparency were prioritized in reporting the methodologies, results, and limitations. Full disclosure of any challenges encountered during the study was provided, ensuring that the conclusions drawn were objective and based on sound and rigorous data analysis. The ANOVA method was used to rigorously evaluate the data. Furthermore, the principle of integrity was upheld by ensuring that the data was not misrepresented and that proper citations were included to acknowledge the contributions of others, thereby preventing plagiarism and recognizing prior work. Since this study relied on secondary data obtained from publicly available datasets on Kaggle, ethical approval from NACOSTI was obtained to ensure compliance with national ethical guidelines. While the research did not involve direct human participants, the ethical approval guaranteed that all procedures related to data collection, storage, data privacy and analysis adhered to established legal and ethical standards. The approval letters from the university and NACOSTI have been appended in the appendices of this thesis. Lastly, since secondary data did not involve the direct interaction with human subjects, ethical considerations, such as respecting data ownership, confidentiality, and acknowledging the source of the datasets, was observed throughout the research. Additionally, data usage adhered to Kaggle's terms of service, ensuring that all legal and ethical conditions associated with the datasets were respected.

While sentiment analysis models offered valuable insights, they posed ethical risks related to bias, privacy, and fairness. These risks included:

i. Algorithmic Bias – Sentiment models often reflect biases present in training data. For instance, reviews of high-end mobile phones may contain subtly positive phrasing that skews classification results. This study incorporated data augmentation techniques to improve model fairness across diverse datasets.

ii. Privacy Concerns – Sentiment analysis involves processing user-generated content, raising concerns about data ownership and consent. This research ensured compliance with data protection frameworks (e.g., GDPR), anonymizing datasets and avoiding personally identifiable information.

iii. Fairness in Sentiment Classification – Implicit negation is linguistically complex, leading to misclassification disparities across demographic groups.

To mitigate this, the model was tested across multiple platforms (Amazon, Google Play, iPhone App Store) to ensure robustness and reduce domain-specific bias. By addressing these concerns, the proposed model aimed to enhance transparency, fairness, and ethical AI deployment in sentiment analysis applications.

3.7 Summary

Chapter three presents the methodology of how the four research objectives stated in chapter one were accomplished. The philosophy and design strategies used in this study are discussed. This research work adopted pragmatism research philosophy that allowed the researcher to utilize the mixed methods research design. In the research design section, the overall strategy that guided the execution of this study is described. The experimental research process involved three key phases: the empirical baseline analysis, the hybrid model development and the model evaluation experiments. Further the main experiments, population sampling strategy, research instruments and data collection process are discussed. Purposeful sampling was adopted in this study in order to select the variables for use in the main experiments of the empirical baseline analysis process. The research instruments utilized were the observation and experimental checklist. The two ensures a rigorous and systematic approach to both the experiments and the data analysis.

The experimental data collection process integrated both the experimental checklists and observation checklists to ensure a systematic approach to deriving performance metrics from various sentiment analysis models. Data analysis utilized in this research study is discussed. This study incorporates both quantitative and qualitative approaches to allow for comprehensive evaluation of the performance of sentiment analysis models. Finally the ethical consideration ensured the hybrid model enhance transparency, fairness, and ethical AI deployment in sentiment analysis applications.

CHAPTER FOUR

EVALUATION OF THE EXISTING SENTIMENT ANALYSIS MODELS FOR DETECTING POLARITY SHIFT

4.1 Introduction

This chapter presents the results of the experiments conducted to evaluate the effectiveness of the current sentiment analysis models and preprocessing techniques, as outlined in Chapter 1. The main objective of this study was to evaluate the performance of current models in detecting polarity shifts caused by implicit negation, with a particular focus on mobile phone brand reviews. The experiments in the empirical baseline analysis were designed to assess the impact of the various sentiment analysis models, feature extraction, feature selection and data augmentation techniques. As stated in Chapter one, the second objective of this research was to evaluate the existing sentiment analysis models, specifically their ability to address polarity shifts due to implicit negation. To achieve this, experiments were set up to test the performance of various models on Amazon Mobile Phone Reviews dataset, in a variety of feature selection, feature extraction and data augmentation techniques. Section 4.2 presents the lessons learnt from the pilot study. Section 4.3 provides a detailed analysis of the experimental data, including the comparison of models and their effectiveness in handling polarity shifts. Sections 4.3.1 to 4.3.8 present the results of ANOVA statistical analyses. Section 4.4 discusses the experimental results. Section 4.5 presents the key findings from the empirical analysis, highlighting how these insights guided the development of the new hybrid model. Lastly a summary of the chapter is done in section 4.6.

4.2 Pilot Study Results

The selection of the ten models in the empirical baseline analysis is grounded in the empirical findings from the pilot study [173]. In line with the systematic review works by Murithi et al [124], a wider range of the traditional machine learning and deep learning models, including individual algorithms and hybrid configurations, were identified for their application in sentiment polarity classification tasks, through a comprehensive theoretical analysis. These choices ensured that the research was both methodologically rigorous and optimally aligned with the research objectives.

Besides this, the preliminary findings in the pilot study also helped in validating or challenging the theoretical expectations from the theoretical analysis.

The challenges identified during the pilot study are used as the basis for improving the experimental design of the empirical baseline process. For example, through the pilot study, it was possible to note that running the experiments on the Google Colab's cloud environment was more computationally efficient as opposed to running them on the Jupyter Notebook on a local machine. The pilot study provided an opportunity to refine both the experimental procedures in the experimental checklists as well as the observation check lists used as part the main research instruments during the empirical analysis process.

During the pilot study, a cross sectional survey on the independent, moderating and dependent variables were investigated in order to refine the experimental research design of the experiments in the empirical analysis process. This exploratory piloting phase revealed that SVM and Naïve Bayes classifiers performed better than logistic regression while the BERT and LSTM performed better than Gated recurrent unit (GRU). It is evidenced from the pilot study that both the support vector machine and the naive Bayesian perform relatively well across most of the datasets as compared to the Logistic regression. Similarly, BERT and LSTM perform equally relatively better across most of the datasets as compared to GRU. This observation in this pilot study underpinned the selection of the support vector machine, naïve Bayesian, LSTM and the BERT models in the empirical baseline analysis. This also formed the basis of the part of sampling strategy for these models in line with the empirical algorithmic study of Gikera et al.[174].

4.3 Experimental Results

Empirical algorithmic involves the empirical analysis of algorithms to understand their behaviour, performance, and efficiency [175]. It focuses on experimentally evaluating algorithms using real-world data rather than solely relying on theoretical analysis [176]. Empirical algorithmics involved conducting experiments on algorithms using various datasets to evaluate their performance under different conditions, comparing their performance on standardized datasets to establish

benchmarks. Through this, it optimizes the algorithmic parameters to enhance performance on specific datasets[175].

Empirical algorithmic bridges the gap between theoretical analysis and real-world applicability by providing empirical evidence of how algorithms behave in practice[176]. Empirical analysis in machine learning is conducted in order to provide a reference for models selection based on their empirical applicability on a real world dataset [177]. In light of this, the experiments in this research were conducted in order to answer the second research question formulated in chapter 1 and aimed at investigating the performance and interactions of the current sentiment analysis models in a variety of feature selection, feature extraction and data augmentation techniques. Table 4.2 is a summary of the set of independent variables, moderating variables and interactions used in the analysis of the different sets of experiments. The scores for the Cohen's kappa, Matthew's correlation and accuracy were used as the dependent variables. In the table 4.2, all the experimental data were analysed using 3-way ANOVA. The three performance scores were first normalized using the Equation 3.1 before being reported as a single average performance score. During normalization, the range was first identified followed by the identification of both the minimum and maximum values. For each data point x, the minimum value was subtracted from it. This formulated the denominator. The difference between the maximum and the minimum values was formulated as the denominator. The numerator divided by the denominator results into the normalized score for a single index.

Table 4.1: Summary of the Set of Variables, Interactions and the Data Analysis Method Applied in the Different Set of Experiments.

Independent variable	Moderating variable	Interaction
Sentiment classification Models	a) Feature Selection Techniques b) Data Augmentation Techniques	Sentiment classification models, Feature extraction and data augmentation techniques
Sentiment classification Models	a) Feature Selection Techniques b) Feature Extraction Methods	Sentiment A classification models, Feature Extraction and Feature Selection techniques.
Sentiment classification Models	a) Data Augmentation Techniques b) Feature Extraction Methods	Sentiment classification models, Feature Extraction and Data Augmentation techniques.
None	a) Feature Selection Techniques b) Data Augmentation Techniques c) Feature Extraction Methods	Feature Extraction, Feature Selection and Data Augmentation techniques

The analysed results for the experimental data from the empirical analysis were presented in sections 4.3.2 to 4.3.9. For the ANOVA analyses, a confidence level of 95% was used. This indicates the level of certainty in the statistical analysis where if the same study is repeated multiple times, the result would fall within the given confidence interval. The sum of squares in the ANOVA tables is a measure of the variation (dispersion) in the experimental data that is attributed to different sources. It evaluated the significance of the techniques, datasets and the dimensionality reduction methods in explaining the variability in the internal validation indexes and the cluster building times. Df (degree of freedom) are associated with different sources of variation in the performance scores of the sentiment analysis models, and are critical for determining the significance of those sources. Mean square is measures of the average variance of the squared differences within a specific source of variation e.g. feature extraction method. It is a key component in the F-ratio calculation in this research study, which was used to determine the statistical significance of the sources of variation.

The F-statistic is the ratio of the variance between groups to the variance within groups. It assesses whether the means of the different sentiment analysis models being compared, feature extraction techniques, feature selection techniques and data augmentation techniques were significantly different. It helps determine whether there are significant differences. The *P*-value is a statistical measure that helps assess the significance of the F-statistic. The *p*-value provides the probability of observing an F-statistic as extreme as the one calculated, assuming that the null hypothesis is true.

In this context, the *P*-value determined the significance of the effect of the sentiment analysis models being compared, feature extraction techniques, feature selection techniques and data augmentation techniques on the performance. In this context, the residuals represented the discrepancies between the actual and predicted performance based on the sentiment analysis models. The normalities of the residuals were tested using the Q-Q plots where the presence of any deviations from the diagonal lines was verified.

4.3.1 One-Way ANOVA Analysis of the Average Normalized Score on Sentiment Classification Model

Table 4.3 shows one-way ANOVA analysis results on the average normalized score of the existing sentiment classification models.

Table 4.2: One-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model

ANOVA - Performance

	Sum of Squares	df	Mean Square	F	p
Model	4.04	9	0.4485	28.9	< .001
Residuals	1.09	70	0.0155		

The low p-value of less than 0.001 on was a clear indication that there were significant variations in the performance scores across the existing sentiment

analysis models used in handling the polarity shift. Based on this evidence, a further post-hoc analysis was performed and presented in tables 4.4 below.

Table 4.3: Post Hoc Analysis of the Average Normalized Score on Sentiment Analysis Model

Post Hoc Comparisons - Model

Model		Model	Mean Difference	SE	df	t	p_{tukey}
DL1	-	DL2	-0.01875	0.0623	70.0	-0.301	1.000
	-	HDL1	-0.04000	0.0623	70.0	-0.642	1.000
	-	HDL2	-0.04000	0.0623	70.0	-0.642	1.000
	-	HTMDL1	0.06000	0.0623	70.0	0.963	0.993
	-	HTMDL2	0.06000	0.0623	70.0	0.963	0.993
	-	HTML1	0.47000	0.0623	70.0	7.547	< .001
	-	HTML2	0.40750	0.0623	70.0	6.543	< .001
	-	TML1	0.48500	0.0623	70.0	7.788	< .001
	-	TML2	0.46000	0.0623	70.0	7.386	< .001
DL2	-	HDL1	-0.02125	0.0623	70.0	-0.341	1.000
	-	HDL2	-0.02125	0.0623	70.0	-0.341	1.000
	-	HTMDL1	0.07875	0.0623	70.0	1.265	0.958
	-	HTMDL2	0.07875	0.0623	70.0	1.265	0.958
	-	HTML1	0.48875	0.0623	70.0	7.848	< .001
	-	HTML2	0.42625	0.0623	70.0	6.844	< .001
	-	TML1	0.50375	0.0623	70.0	8.089	< .001
	-	TML2	0.47875	0.0623	70.0	7.688	< .001
HDL1	-	HDL2	2.78e-16	0.0623	70.0	4.46e-15	1.000
	-	HTMDL1	0.10000	0.0623	70.0	1.606	0.841
	-	HTMDL2	0.10000	0.0623	70.0	1.606	0.841
	-	HTML1	0.51000	0.0623	70.0	8.189	< .001
	-	HTML2	0.44750	0.0623	70.0	7.186	< .001
	-	TML1	0.52500	0.0623	70.0	8.430	< .001
	-	TML2	0.50000	0.0623	70.0	8.029	< .001
HDL2	-	HTMDL1	0.10000	0.0623	70.0	1.606	0.841
	-	HTMDL2	0.10000	0.0623	70.0	1.606	0.841
	-	HTML1	0.51000	0.0623	70.0	8.189	< .001
	-	HTML2	0.44750	0.0623	70.0	7.186	< .001

	-	TML1	0.52500	0.0623	70.0	8.430	< .001
	-	TML2	0.50000	0.0623	70.0	8.029	< .001
HTMDL1	-	HTMDL2	4.72e-16	0.0623	70.0	7.58e-15	1.000
	-	HTML1	0.41000	0.0623	70.0	6.584	< .001
	-	HTML2	0.34750	0.0623	70.0	5.580	< .001
	-	TML1	0.42500	0.0623	70.0	6.824	< .001
	-	TML2	0.40000	0.0623	70.0	6.423	< .001
HTMDL2	-	HTML1	0.41000	0.0623	70.0	6.584	< .001
	-	HTML2	0.34750	0.0623	70.0	5.580	< .001
	-	TML1	0.42500	0.0623	70.0	6.824	< .001
	-	TML2	0.40000	0.0623	70.0	6.423	< .001
HTML1	-	HTML2	-0.06250	0.0623	70.0	-1.004	0.991
	-	TML1	0.01500	0.0623	70.0	0.241	1.000
	-	TML2	-0.01000	0.0623	70.0	-0.161	1.000
HTML2	-	TML1	0.07750	0.0623	70.0	1.244	0.962
	-	TML2	0.05250	0.0623	70.0	0.843	0.998
TML1	-	TML2	-0.02500	0.0623	70.0	-0.401	1.000

Note. Comparisons are based on estimated marginal means

From the post-hoc analysis, it was clear that there was a statistic significant difference noticeable among the performance of deep learning and traditional models. There was also a notable difference between hybrid of deep learning and hybrid traditional machine learning model.

4.3.2 Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Extraction Techniques

The table 4.4 shows the results of the two-way ANOVA analysis on the average normalized score on sentiment analysis model in a variety of feature extraction techniques.

Table 4.4:Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Extraction Techniques

ANOVA - Performance

	Sum of Squares	df	Mean Square	F	p
Model	4.037	9	0.4485	35.32	< .001
Feature Extraction	0.109	1	0.1088	8.57	0.005
Model ✻ Feature Extraction	0.215	9	0.0239	1.88	0.072
Residuals	0.762	60	0.0127		

The low p-values of less than 0.001 on the sentiment analysis models was a clear indication that there are statistically significant variations in the existing models in different feature extraction techniques. Based on this evidence, a further post-hoc analysis was performed and presented in table 4.5.

Table 4.5: Post Hoc Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Extraction Techniques

Post Hoc Comparisons - Model ✻ Feature Extraction

Model	Feature Extraction		Model	Feature Extraction	Mean Difference	SE	df	t	p_{tukey}
DL1	FE1	-	DL1	FE2	-8.33e−17	0.0797	60.0	-1.05e−15	1.000
		-	DL2	FE1	-4.37e−16	0.0797	60.0	-5.49e−15	1.000
		-	DL2	FE2	-0.03750	0.0797	60.0	-0.4706	1.000
		-	HDL1	FE1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL1	FE2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FE1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FE2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HTMDL1	FE1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL1	FE2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FE1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FE2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTML1	FE1	0.39500	0.0797	60.0	4.9573	< .001
		-	HTML1	FE2	0.54500	0.0797	60.0	6.8398	< .001

Post Hoc Comparisons - Model ✻ Feature Extraction

Model	Feature Extraction		Model	Feature Extraction	Mean Difference	SE	df	t	p_{tukey}
		-	HTML2	FE1	0.27000	0.0797	60.0	3.3885	0.113
		-	HTML2	FE2	0.54500	0.0797	60.0	6.8398	< .001
		-	TML1	FE1	0.41000	0.0797	60.0	5.1456	< .001
		-	TML1	FE2	0.56000	0.0797	60.0	7.0281	< .001
		-	TML2	FE1	0.36000	0.0797	60.0	4.5181	0.004
		-	TML2	FE2	0.56000	0.0797	60.0	7.0281	< .001
	FE2	-	DL2	FE1	-3.54e−16	0.0797	60.0	-4.44e−15	1.000
		-	DL2	FE2	-0.03750	0.0797	60.0	-0.4706	1.000
		-	HDL1	FE1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL1	FE2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FE1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FE2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HTMDL1	FE1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL1	FE2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FE1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FE2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTML1	FE1	0.39500	0.0797	60.0	4.9573	< .001
		-	HTML1	FE2	0.54500	0.0797	60.0	6.8398	< .001
		-	HTML2	FE1	0.27000	0.0797	60.0	3.3885	0.113
		-	HTML2	FE2	0.54500	0.0797	60.0	6.8398	< .001
		-	TML1	FE1	0.41000	0.0797	60.0	5.1456	< .001
		-	TML1	FE2	0.56000	0.0797	60.0	7.0281	< .001
		-	TML2	FE1	0.36000	0.0797	60.0	4.5181	0.004
		-	TML2	FE2	0.56000	0.0797	60.0	7.0281	< .001
DL2	FE1	-	DL2	FE2	-0.03750	0.0797	60.0	-0.4706	1.000
		-	HDL1	FE1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL1	FE2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FE1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FE2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HTMDL1	FE1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL1	FE2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FE1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FE2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTML1	FE1	0.39500	0.0797	60.0	4.9573	< .001
		-	HTML1	FE2	0.54500	0.0797	60.0	6.8398	< .001
		-	HTML2	FE1	0.27000	0.0797	60.0	3.3885	0.113
		-	HTML2	FE2	0.54500	0.0797	60.0	6.8398	< .001
		-	TML1	FE1	0.41000	0.0797	60.0	5.1456	< .001
		-	TML1	FE2	0.56000	0.0797	60.0	7.0281	< .001
		-	TML2	FE1	0.36000	0.0797	60.0	4.5181	0.004

Post Hoc Comparisons - Model ✻ Feature Extraction

Model	Feature Extraction		Model	Feature Extraction	Mean Difference	SE	df	t	p_{tukey}
		-	TML2	FE2	0.56000	0.0797	60.0	7.0281	< .001
	FE2	-	HDL1	FE1	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HDL1	FE2	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HDL2	FE1	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HDL2	FE2	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HTMDL1	FE1	0.09750	0.0797	60.0	1.2236	0.999
		-	HTMDL1	FE2	0.09750	0.0797	60.0	1.2236	0.999
		-	HTMDL2	FE1	0.09750	0.0797	60.0	1.2236	0.999
		-	HTMDL2	FE2	0.09750	0.0797	60.0	1.2236	0.999
		-	HTML1	FE1	0.43250	0.0797	60.0	5.4279	< .001
		-	HTML1	FE2	0.58250	0.0797	60.0	7.3105	< .001
		-	HTML2	FE1	0.30750	0.0797	60.0	3.8592	0.033
		-	HTML2	FE2	0.58250	0.0797	60.0	7.3105	< .001
		-	TML1	FE1	0.44750	0.0797	60.0	5.6162	< .001
		-	TML1	FE2	0.59750	0.0797	60.0	7.4987	< .001
		-	TML2	FE1	0.39750	0.0797	60.0	4.9887	< .001
		-	TML2	FE2	0.59750	0.0797	60.0	7.4987	< .001
HDL1	FE1	-	HDL1	FE2	-5.55e−17	0.0797	60.0	-6.97e−16	1.000
		-	HDL2	FE1	2.57e-16	0.0797	60.0	3.22e-15	1.000
		-	HDL2	FE2	2.43e-16	0.0797	60.0	3.05e-15	1.000
		-	HTMDL1	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FE1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	HTML2	FE1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FE1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FE2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FE1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FE2	0.60000	0.0797	60.0	7.5301	< .001
	FE2	-	HDL2	FE1	3.12e-16	0.0797	60.0	3.92e-15	1.000
		-	HDL2	FE2	2.98e-16	0.0797	60.0	3.74e-15	1.000
		-	HTMDL1	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FE1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FE2	0.58500	0.0797	60.0	7.3418	< .001

Post Hoc Comparisons - Model ✻ Feature Extraction

Model	Feature Extraction		Model	Feature Extraction	Mean Difference	SE	df	t	p_{tukey}
		-	HTML2	FE1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FE1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FE2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FE1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FE2	0.60000	0.0797	60.0	7.5301	< .001
HDL2	FE1	-	HDL2	FE2	-1.39e−17	0.0797	60.0	-1.74e−16	1.000
		-	HTMDL1	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FE1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	HTML2	FE1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FE1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FE2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FE1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FE2	0.60000	0.0797	60.0	7.5301	< .001
	FE2	-	HTMDL1	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FE2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FE1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	HTML2	FE1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FE2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FE1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FE2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FE1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FE2	0.60000	0.0797	60.0	7.5301	< .001
HTMDL1	FE1	-	HTMDL1	FE2	-2.78e−17	0.0797	60.0	-3.48e−16	1.000
		-	HTMDL2	FE1	5.00e-16	0.0797	60.0	6.27e-15	1.000
		-	HTMDL2	FE2	4.44e-16	0.0797	60.0	5.57e-15	1.000
		-	HTML1	FE1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FE1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FE1	0.35000	0.0797	60.0	4.3926	0.006

Post Hoc Comparisons - Model ✻ Feature Extraction

Model	Feature Extraction		Model	Feature Extraction	Mean Difference	SE	df	t	p_{tukey}
		-	TML1	FE2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FE1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FE2	0.50000	0.0797	60.0	6.2751	< .001
	FE2	-	HTMDL2	FE1	5.27e-16	0.0797	60.0	6.62e-15	1.000
		-	HTMDL2	FE2	4.72e-16	0.0797	60.0	5.92e-15	1.000
		-	HTML1	FE1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FE1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FE1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FE2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FE1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FE2	0.50000	0.0797	60.0	6.2751	< .001
HTMDL2	FE1	-	HTMDL2	FE2	-5.55e−17	0.0797	60.0	-6.97e−16	1.000
		-	HTML1	FE1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FE1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FE1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FE2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FE1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FE2	0.50000	0.0797	60.0	6.2751	< .001
	FE2	-	HTML1	FE1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FE1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FE2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FE1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FE2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FE1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FE2	0.50000	0.0797	60.0	6.2751	< .001
HTML1	FE1	-	HTML1	FE2	0.15000	0.0797	60.0	1.8825	0.932
		-	HTML2	FE1	-0.12500	0.0797	60.0	-1.5688	0.988
		-	HTML2	FE2	0.15000	0.0797	60.0	1.8825	0.932
		-	TML1	FE1	0.01500	0.0797	60.0	0.1883	1.000
		-	TML1	FE2	0.16500	0.0797	60.0	2.0708	0.858
		-	TML2	FE1	-0.03500	0.0797	60.0	-0.4393	1.000
		-	TML2	FE2	0.16500	0.0797	60.0	2.0708	0.858
	FE2	-	HTML2	FE1	-0.27500	0.0797	60.0	-3.4513	0.097
		-	HTML2	FE2	5.41e-16	0.0797	60.0	6.79e-15	1.000
		-	TML1	FE1	-0.13500	0.0797	60.0	-1.6943	0.974

Post Hoc Comparisons - Model ✻ Feature Extraction

Model	Feature Extraction		Model	Feature Extraction	Mean Difference	SE	df	t	p_{tukey}
		-	TML1	FE2	0.01500	0.0797	60.0	0.1883	1.000
		-	TML2	FE1	-0.18500	0.0797	60.0	-2.3218	0.712
		-	TML2	FE2	0.01500	0.0797	60.0	0.1883	1.000
HTML2	FE1	-	HTML2	FE2	0.27500	0.0797	60.0	3.4513	0.097
		-	TML1	FE1	0.14000	0.0797	60.0	1.7570	0.963
		-	TML1	FE2	0.29000	0.0797	60.0	3.6395	0.060
		-	TML2	FE1	0.09000	0.0797	60.0	1.1295	1.000
		-	TML2	FE2	0.29000	0.0797	60.0	3.6395	0.060
	FE2	-	TML1	FE1	-0.13500	0.0797	60.0	-1.6943	0.974
		-	TML1	FE2	0.01500	0.0797	60.0	0.1883	1.000
		-	TML2	FE1	-0.18500	0.0797	60.0	-2.3218	0.712
		-	TML2	FE2	0.01500	0.0797	60.0	0.1883	1.000
TML1	FE1	-	TML1	FE2	0.15000	0.0797	60.0	1.8825	0.932
		-	TML2	FE1	-0.05000	0.0797	60.0	-0.6275	1.000
		-	TML2	FE2	0.15000	0.0797	60.0	1.8825	0.932
	FE2	-	TML2	FE1	-0.20000	0.0797	60.0	-2.5100	0.581
		-	TML2	FE2	3.61e-16	0.0797	60.0	4.53e-15	1.000
TML2	FE1	-	TML2	FE2	0.20000	0.0797	60.0	2.5100	0.581

Note. Comparisons are based on estimated marginal means

From the post hoc analysis, it was clear that there were specific variations in the sentiment analysis models and the feature extraction techniques. The main variations were between hybrid of traditional machine learning models and feature extraction as well among the traditional machine learning and feature extraction techniques.

4.3.3 Two-way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Selection Techniques

The table 4.6 shows the results of the two-way ANOVA analysis on the average normalized score of the existing sentiment analysis models in a variety of feature selection techniques.

Table 4.6: Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Feature Selection Techniques

ANOVA - Performance

	Sum of Squares	df	Mean Square	F	p
Model	4.037	9	0.4485	35.32	< .001
Feature Selection	0.109	1	0.1088	8.57	0.005
Model ✻ Feature Selection	0.215	9	0.0239	1.88	0.072
Residuals	0.762	60	0.0127		

Two-way ANOVA analysis was conducted to examine the effects of feature selection technique on the sentiment analysis models. It aimed to investigate whether there is an interaction between the sentiment analysis model and Feature Selection Technique, meaning whether their combined effect differs from their individual effects. The dependent variable in this analysis is the average normalized score, representing the model's classification performance. The *P*-value <0.01 was a clear indication that there was interaction between the sentiment analysis and feature selection techniques. Based on this evidence, the post-hoc analysis was performed and the results presented in table 4.7.

Table 4.7: Post Hoc Analysis of Sentiment Analysis Models and Feature Selection Techniques

Table Post Hoc Comparisons - Model ✻ Feature Selection

Model	Feature Selection		Model	Feature Selection	Mean Difference	SE	df	t	ptukey
DL1	FS1	-	DL1	FS2	-2.08e−16	0.0797	60.0	-2.61e−15	1.000
		-	DL2	FS1	-4.44e−16	0.0797	60.0	-5.57e−15	1.000
		-	DL2	FS2	-0.03750	0.0797	60.0	-0.4706	1.000
		-	HDL1	FS1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL1	FS2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FS1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FS2	-0.04000	0.079	60.	-0.5020	1.000

					7	0		
	-	HTMDL1	FS1	0.06000	0.0797	60.0	0.7530	1.000
	-	HTMDL1	FS2	0.06000	0.0797	60.0	0.7530	1.000
	-	HTMDL2	FS1	0.06000	0.0797	60.0	0.7530	1.000
	-	HTMDL2	FS2	0.06000	0.0797	60.0	0.7530	1.000
	-	HTML1	FS1	0.39500	0.0797	60.0	4.9573	< .001
	-	HTML1	FS2	0.54500	0.0797	60.0	6.8398	< .001
	-	HTML2	FS1	0.27000	0.0797	60.0	3.3885	0.113
	-	HTML2	FS2	0.54500	0.0797	60.0	6.8398	< .001
	-	TML1	FS1	0.41000	0.0797	60.0	5.1456	< .001
	-	TML1	FS2	0.56000	0.0797	60.0	7.0281	< .001
	-	TML2	FS1	0.36000	0.0797	60.0	4.5181	0.004
	-	TML2	FS2	0.56000	0.0797	60.0	7.0281	< .001
FS2	-	DL2	FS1	-2.36e−16	0.0797	60.0	-2.96e−15	1.000
	-	DL2	FS2	-0.03750	0.0797	60.0	-0.4706	1.000
	-	HDL1	FS1	-0.04000	0.0797	60.0	-0.5020	1.000
	-	HDL1	FS2	-0.04000	0.0797	60.0	-0.5020	1.000
	-	HDL2	FS1	-0.04000	0.0797	60.0	-0.5020	1.000
	-	HDL2	FS2	-0.04000	0.0797	60.0	-0.5020	1.000
	-	HTMDL1	FS1	0.06000	0.0797	60.0	0.7530	1.000
	-	HTMDL1	FS2	0.06000	0.0797	60.0	0.7530	1.000
	-	HTMDL2	FS1	0.06000	0.0797	60.0	0.7530	1.000
	-	HTMDL2	FS2	0.06000	0.0797	60.0	0.7530	1.000
	-	HTML1	FS1	0.39500	0.0797	60.0	4.9573	< .001
	-	HTML1	FS2	0.54500	0.0797	60.0	6.8398	< .001

		-	HTML2	FS1	0.27000	0.0797	60.0	3.3885	0.113
		-	HTML2	FS2	0.54500	0.0797	60.0	6.8398	< .001
		-	TML1	FS1	0.41000	0.0797	60.0	5.1456	< .001
		-	TML1	FS2	0.56000	0.0797	60.0	7.0281	< .001
		-	TML2	FS1	0.36000	0.0797	60.0	4.5181	0.004
		-	TML2	FS2	0.56000	0.0797	60.0	7.0281	< .001
DL2	FS1	-	DL2	FS2	-0.03750	0.0797	60.0	-0.4706	1.000
		-	HDL1	FS1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL1	FS2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FS1	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HDL2	FS2	-0.04000	0.0797	60.0	-0.5020	1.000
		-	HTMDL1	FS1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL1	FS2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FS1	0.06000	0.0797	60.0	0.7530	1.000
		-	HTMDL2	FS2	0.06000	0.0797	60.0	0.7530	1.000
		-	HTML1	FS1	0.39500	0.0797	60.0	4.9573	< .001
		-	HTML1	FS2	0.54500	0.0797	60.0	6.8398	< .001
		-	HTML2	FS1	0.27000	0.0797	60.0	3.3885	0.113
		-	HTML2	FS2	0.54500	0.0797	60.0	6.8398	< .001
		-	TML1	FS1	0.41000	0.0797	60.0	5.1456	< .001
		-	TML1	FS2	0.56000	0.0797	60.0	7.0281	< .001
		-	TML2	FS1	0.36000	0.0797	60.0	4.5181	0.004
		-	TML2	FS2	0.56000	0.0797	60.0	7.0281	< .001
	FS2	-	HDL1	FS1	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HDL1	FS2	-0.00250	0.0797	60.0	-0.0314	1.000

		-	HDL2	FS1	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HDL2	FS2	-0.00250	0.0797	60.0	-0.0314	1.000
		-	HTMDL1	FS1	0.09750	0.0797	60.0	1.2236	0.999
		-	HTMDL1	FS2	0.09750	0.0797	60.0	1.2236	0.999
		-	HTMDL2	FS1	0.09750	0.0797	60.0	1.2236	0.999
		-	HTMDL2	FS2	0.09750	0.0797	60.0	1.2236	0.999
		-	HTML1	FS1	0.43250	0.0797	60.0	5.4279	< .001
		-	HTML1	FS2	0.58250	0.0797	60.0	7.3105	< .001
		-	HTML2	FS1	0.30750	0.0797	60.0	3.8592	0.033
		-	HTML2	FS2	0.58250	0.0797	60.0	7.3105	< .001
		-	TML1	FS1	0.44750	0.0797	60.0	5.6162	< .001
		-	TML1	FS2	0.59750	0.0797	60.0	7.4987	< .001
		-	TML2	FS1	0.39750	0.0797	60.0	4.9887	< .001
		-	TML2	FS2	0.59750	0.0797	60.0	7.4987	< .001
HDL1	FS1	-	HDL1	FS2	-8.33e−17	0.0797	60.0	-1.05e−15	1.000
		-	HDL2	FS1	2.29e-16	0.0797	60.0	2.87e-15	1.000
		-	HDL2	FS2	2.43e-16	0.0797	60.0	3.05e-15	1.000
		-	HTMDL1	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FS1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FS2	0.58500	0.0797	60.0	7.3418	< .001
		-	HTML2	FS1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FS2	0.58500	0.0797	60.0	7.3418	< .001

		-	TML1	FS1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FS2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FS1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FS2	0.60000	0.0797	60.0	7.5301	< .001
	FS2	-	HDL2	FS1	3.12e-16	0.0797	60.0	3.92e-15	1.000
		-	HDL2	FS2	3.26e-16	0.0797	60.0	4.09e-15	1.000
		-	HTMDL1	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FS1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FS2	0.58500	0.0797	60.0	7.3418	< .001
		-	HTML2	FS1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FS2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FS1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FS2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FS1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FS2	0.60000	0.0797	60.0	7.5301	< .001
HDL2	FS1	-	HDL2	FS2	1.39e-17	0.0797	60.0	1.74e-16	1.000
		-	HTMDL1	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FS1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FS2	0.58500	0.0797	60.0	7.3418	< .001

		-	HTML2	FS1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FS2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FS1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FS2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FS1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FS2	0.60000	0.0797	60.0	7.5301	< .001
	FS2	-	HTMDL1	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL1	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS1	0.10000	0.0797	60.0	1.2550	0.999
		-	HTMDL2	FS2	0.10000	0.0797	60.0	1.2550	0.999
		-	HTML1	FS1	0.43500	0.0797	60.0	5.4593	< .001
		-	HTML1	FS2	0.58500	0.0797	60.0	7.3418	< .001
		-	HTML2	FS1	0.31000	0.0797	60.0	3.8905	0.030
		-	HTML2	FS2	0.58500	0.0797	60.0	7.3418	< .001
		-	TML1	FS1	0.45000	0.0797	60.0	5.6476	< .001
		-	TML1	FS2	0.60000	0.0797	60.0	7.5301	< .001
		-	TML2	FS1	0.40000	0.0797	60.0	5.0201	< .001
		-	TML2	FS2	0.60000	0.0797	60.0	7.5301	< .001
HTMDL1	FS1	-	HTMDL1	FS2	5.55e-17	0.0797	60.0	6.97e-16	1.000
		-	HTMDL2	FS1	5.07e-16	0.0797	60.0	6.36e-15	1.000
		-	HTMDL2	FS2	5.20e-16	0.0797	60.0	6.53e-15	1.000
		-	HTML1	FS1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FS2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FS1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FS2	0.48500	0.0797	60.0	6.0868	< .001

		-	TML1	FS1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FS2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FS1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FS2	0.50000	0.0797	60.0	6.2751	< .001
	FS2	-	HTMDL2	FS1	4.51e-16	0.0797	60.0	5.66e-15	1.000
		-	HTMDL2	FS2	4.65e-16	0.0797	60.0	5.83e-15	1.000
		-	HTML1	FS1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FS2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FS1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FS2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FS1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FS2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FS1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FS2	0.50000	0.0797	60.0	6.2751	< .001
HTMDL2	FS1	-	HTMDL2	FS2	1.39e-17	0.0797	60.0	1.74e-16	1.000
		-	HTML1	FS1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FS2	0.48500	0.0797	60.0	6.0868	< .001
		-	HTML2	FS1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FS2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FS1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FS2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FS1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FS2	0.50000	0.0797	60.0	6.2751	< .001
	FS2	-	HTML1	FS1	0.33500	0.0797	60.0	4.2043	0.012
		-	HTML1	FS2	0.48500	0.0797	60.0	6.0868	< .001

		-	HTML2	FS1	0.21000	0.0797	60.0	2.6355	0.492
		-	HTML2	FS2	0.48500	0.0797	60.0	6.0868	< .001
		-	TML1	FS1	0.35000	0.0797	60.0	4.3926	0.006
		-	TML1	FS2	0.50000	0.0797	60.0	6.2751	< .001
		-	TML2	FS1	0.30000	0.0797	60.0	3.7650	0.043
		-	TML2	FS2	0.50000	0.0797	60.0	6.2751	< .001
HTML1	FS1	-	HTML1	FS2	0.15000	0.0797	60.0	1.8825	0.932
		-	HTML2	FS1	-0.12500	0.0797	60.0	-1.5688	0.988
		-	HTML2	FS2	0.15000	0.0797	60.0	1.8825	0.932
		-	TML1	FS1	0.01500	0.0797	60.0	0.1883	1.000
		-	TML1	FS2	0.16500	0.0797	60.0	2.0708	0.858
		-	TML2	FS1	-0.03500	0.0797	60.0	-0.4393	1.000
		-	TML2	FS2	0.16500	0.0797	60.0	2.0708	0.858
	FS2	-	HTML2	FS1	-0.27500	0.0797	60.0	-3.4513	0.097
		-	HTML2	FS2	5.48e-16	0.0797	60.0	6.88e-15	1.000
		-	TML1	FS1	-0.13500	0.0797	60.0	-1.6943	0.974
		-	TML1	FS2	0.01500	0.0797	60.0	0.1883	1.000
		-	TML2	FS1	-0.18500	0.0797	60.0	-2.3218	0.712
		-	TML2	FS2	0.01500	0.0797	60.0	0.1883	1.000
HTML2	FS1	-	HTML2	FS2	0.27500	0.0797	60.0	3.4513	0.097
		-	TML1	FS1	0.14000	0.0797	60.0	1.7570	0.963
		-	TML1	FS2	0.29000	0.0797	60.0	3.6395	0.060
		-	TML2	FS1	0.09000	0.0797	60.0	1.1295	1.000
		-	TML2	FS2	0.29000	0.0797	60.0	3.6395	0.060
	FS2	-	TML1	FS1	-0.13500	0.0797	60.0	-1.6943	0.974

		-	TML1	FS2	0.01500	0.0797	60.0	0.1883	1.000
		-	TML2	FS1	-0.18500	0.0797	60.0	-2.3218	0.712
		-	TML2	FS2	0.01500	0.0797	60.0	0.1883	1.000
TML1	FS1	-	TML1	FS2	0.15000	0.0797	60.0	1.8825	0.932
		-	TML2	FS1	-0.05000	0.0797	60.0	-0.6275	1.000
		-	TML2	FS2	0.15000	0.0797	60.0	1.8825	0.932
	FS2	-	TML2	FS1	-0.20000	0.0797	60.0	-2.5100	0.581
		-	TML2	FS2	3.68e-16	0.0797	60.0	4.62e-15	1.000
TML2	FS1	-	TML2	FS2	0.20000	0.0797	60.0	2.5100	0.581

Note. Comparisons are based on estimated marginal means

From the post hoc analysis, it was evident that the significant variations in sentiment analysis models and feature selection techniques were only between the traditional machine learning models and their hybrid. Feature selection had no significance interactions with the deep learning models. This was evidenced by the *P*-value of 1 and 0.999. This was clear that choosing the right feature selection techniques for traditional machine learning models and their hybrids was very key for the model's performance.

4.3.4 Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Data Augmentation

The table 4.8 shows the results of the two-way ANOVA analysis of the average normalized score on the sentiment analysis models in a variety of the data augmentation techniques.

Table 4.8:Two-Way ANOVA Analysis of the Average Normalized Score on Sentiment Analysis Model and Data Augmentation

ANOVA - Performance

	Sum of Squares	df	Mean Square	F	p
Model	4.037	9	0.4485	38.49	< .001
Data Augmentation	0.205	1	0.2050	17.59	< .001
Model ✻ Data Augmentation	0.182	9	0.0202	1.73	0.101
Residuals	0.699	60	0.0117		

The results show that there were significant differences between the different data augmentation techniques and the sentiment analysis models. Based on this evidence, the post hoc analysis was performed and reported in table 4.9.

Table 4.9: Post Hoc Analysis of Model and Data augmentation techniques

Model	DA	Model	DA	Mean Difference	SE	df	t	ptukey
TML1	DA1	- TML1	DA2	-1.11e−16	0.0678	40.0	-1.64e−15	1.000
		- TML2	DA1	-1.71e−15	0.0678	40.0	-2.52e−14	1.000
		- TML2	DA2	-0.05000	0.0678	40.0	-0.738	1.000
		- DL1	DA1	-0.57000	0.0678	40.0	-8.412	< .001
		- DL1	DA2	-0.40000	0.0678	40.0	-5.903	< .001
		- DL2	DA1	-0.57000	0.0678	40.0	-8.412	< .001
		- DL2	DA2	-0.43750	0.0678	40.0	-6.456	< .001
		- HTML1	DA1	-0.01500	0.0678	40.0	-0.221	1.000
		- HTML1	DA2	-0.01500	0.0678	40.0	-0.221	1.000
		- HTML2	DA1	-0.07750	0.0678	40.0	-1.144	1.000
		- HTML2	DA2	-0.07750	0.0678	40.0	-1.144	1.000
		- HDL1	DA1	-0.62000	0.0678	40.0	-9.149	< .001
		- HDL1	DA2	-0.43000	0.0678	40.0	-6.346	< .001
		- HDL2	DA1	-0.62000	0.0678	40.0	-9.149	< .001
		- HDL2	DA2	-0.43000	0.0678	40.0	-6.346	< .001
		- HTMDL1	DA1	-0.52000	0.0678	40.0	-7.674	< .001
		- HTMDL1	DA2	-0.33000	0.0678	40.0	-4.870	0.002
		- HTMDL2	DA1	-0.52000	0.0678	40.0	-7.674	< .001
		- HTMDL2	DA2	-0.33000	0.0678	40.0	-4.870	0.002
	DA2	- TML2	DA1	-1.60e−15	0.0678	40.0	-2.36e−14	1.000
		- TML2	DA2	-0.05000	0.0678	40.0	-0.738	1.000
		- DL1	DA1	-0.57000	0.0678	40.0	-8.412	< .001
		- DL1	DA2	-0.40000	0.0678	40.0	-5.903	< .001
		- DL2	DA1	-0.57000	0.0678	40.0	-8.412	< .001

		- DL2	DA2	-0.43750	0.0678	40.0	-6.456	< .001
		- HTML1	DA1	-0.01500	0.0678	40.0	-0.221	1.000
		- HTML1	DA2	-0.01500	0.0678	40.0	-0.221	1.000
		- HTML2	DA1	-0.07750	0.0678	40.0	-1.144	1.000
		- HTML2	DA2	-0.07750	0.0678	40.0	-1.144	1.000
		- HDL1	DA1	-0.62000	0.0678	40.0	-9.149	< .001
		- HDL1	DA2	-0.43000	0.0678	40.0	-6.346	< .001
		- HDL2	DA1	-0.62000	0.0678	40.0	-9.149	< .001
		- HDL2	DA2	-0.43000	0.0678	40.0	-6.346	< .001
		- HTMDL1	DA1	-0.52000	0.0678	40.0	-7.674	< .001
		- HTMDL1	DA2	-0.33000	0.0678	40.0	-4.870	0.002
		- HTMDL2	DA1	-0.52000	0.0678	40.0	-7.674	< .001
		- HTMDL2	DA2	-0.33000	0.0678	40.0	-4.870	0.002
TML2	DA1	- TML2	DA2	-0.05000	0.0678	40.0	-0.738	1.000
		- DL1	DA1	-0.57000	0.0678	40.0	-8.412	< .001
		- DL1	DA2	-0.40000	0.0678	40.0	-5.903	< .001
		- DL2	DA1	-0.57000	0.0678	40.0	-8.412	< .001
		- DL2	DA2	-0.43750	0.0678	40.0	-6.456	< .001
		- HTML1	DA1	-0.01500	0.0678	40.0	-0.221	1.000
		- HTML1	DA2	-0.01500	0.0678	40.0	-0.221	1.000
		- HTML2	DA1	-0.07750	0.0678	40.0	-1.144	1.000

Based on the post-hoc analysis on the sentiment analysis models, it was evident that there are interactions between deep learning models and data augmentation techniques with a *P*-value of less than 0.01. It also applied to hybrid of deep learning models and data augmentation techniques with a *P*-value of less than 0.02.

4.3.5 Three-Way ANOVA Analysis of the Average Normalized Score for Sentiment Analysis Model, Feature Extraction Method and Feature Selection Techniques

The table 4.10 shows the results of the three-way ANOVA analysis of the average normalized score on the sentiment analysis models in a variety of the feature extraction and feature selection techniques.

Table 4.10: Three-Way ANOVA of the Sentiment Analysis Models, Feature Extraction Techniques and the Feature Selection Techniques

ANOVA –Average Normalized Performance

	Sum of Squares	df	Mean Square	F	p
Model	4.03673	9	0.44853	43.743	< .001
Feature Extraction	0.10878	1	0.10878	10.609	0.002
Feature Selection	0.10878	1	0.10878	10.609	0.002
Model ✻ Feature Extraction	0.21528	9	0.02392	2.333	0.032
Model ✻ Feature Selection	0.21528	9	0.02392	2.333	0.032
Feature Extraction ✻ Feature Selection	0.00465	1	0.00465	0.454	0.504
Model ✻ Feature Extraction ✻ Feature Selection	0.02301	9	0.00256	0.249	0.984
Residuals	0.41015	40	0.01025		

The low p-value of less than 0.001 was an indication that there was a statistically significant difference on the performance of the different sentiment analysis models. The low p-value of 0.02 on the feature extraction and feature selection was an indication that they were both a statistically significant factor in the performance of the sentiment analysis models. The p-value of 0.032 indicated there were statistically significant interactions between the sentiment analysis models and feature extraction techniques as well as between sentiment analysis models and the feature selection techniques.

4.3.6 Three-Way ANOVA Analysis of the Average Normalized Score for Sentiment Analysis Model, Feature Selection Method and Data Augmentation Method

Table 4.11 shows three-Way ANOVA conducted to examine the effects of sentiment analysis model, feature selection, and data augmentation technique on the average normalized score.

Table 4.11: Three-Way ANOVA of the Sentiment Analysis Models, Feature Selection Techniques and the Data Augmentation Techniques

ANOVA - Performance Score

	Sum of Squares	df	Mean Square	F	p
Model	4.03673	9	0.44853	48.83906	< .001
Feature Selection	0.10878	1	0.10878	11.84497	0.001
Data Augmentation	0.20503	1	0.20503	22.32544	< .001
Model ✻ Feature Selection	0.21528	9	0.02392	2.60462	0.018
Model ✻ Data Augmentation	0.18168	9	0.02019	2.19810	0.043
Feature Selection ✻ Data Augmentation	3.12e-5	1	3.12e-5	0.00340	0.954
Model ✻ Feature Selection ✻ Data Augmentation	0.00778	9	8.65e-4	0.09414	1.000
Residuals	0.36735	40	0.00918		

The above table exhibits significant interaction between the sentiment analysis models, feature selection techniques as well as the data augmentation techniques, all at their individual levels. The *p*-value of 0.018 shows significant interaction between the sentiment analysis models and the feature selection. There is also a statistical significant interaction between the sentiment analysis model and the data augmentation techniques which is evidenced by the *p*-value of 0.043. Moreover, there are significant interactions between model and data augmentation techniques as well as between models and feature selection techniques. Further post-hoc analysis demonstrates that the significant interactions between the models and the feature selection techniques only apply to the traditional machine learning models only. These interactions do not apply to the deep learning models.

On the other hand, further post-hoc analysis demonstrates that the significant interactions between the sentiment analysis models and data augmentation techniques only apply to the deep learning models. These interactions do not apply to the traditional machine learning models.

4.3.7 Three-Way ANOVA Analysis of the Average Normalized Score for Feature Extraction Method, Feature Selection Method and Data Augmentation Method

Table 4.12 below shows the three-way ANOVA analysis of the average normalized score for the feature extraction, feature selection and data augmentation techniques.

Table 4.12: Three-Way ANOVA of the Feature Extraction Techniques, Feature Selection Techniques and the Data Augmentation Techniques

ANOVA - Average Normalized Performance

	Sum of Squares	df	Mean Square	F	P
Feature Extraction	0.10878	1	0.10878	1.6686	0.201
Data Augmentation	0.20503	1	0.20503	3.1450	0.080
Feature Selection	0.10878	1	0.10878	1.6686	0.201
Feature Extraction ✻ Data Augmentation	3.13e-5	1	3.13e-5	4.79e-4	0.983
Feature Extraction ✻ Feature Selection	0.00465	1	0.00465	0.0713	0.790
Data Augmentation ✻ Feature Selection	3.13e-5	1	3.13e-5	4.79e-4	0.983
Feature Extraction ✻Data Augmentation ✻ Feature Selection	0.00153	1	0.00153	0.0235	0.879
Residuals	4.69383	72	0.06519		

The above table exhibits a lack of significant differences in the feature extraction techniques, data augmentation techniques as well as the feature selection techniques, all at their individual levels. Moreover, there lacks significant interactions between feature extraction and data augmentation techniques, feature extraction and feature selection techniques as well as data augmentation and feature selection techniques. There is also no significant interactions among the three variables combined together i.e. feature extraction techniques, data augmentation techniques and feature selection techniques. This is evidenced by the *p*-value greater than 0.05 for all the results as demonstrated in table 4.12.

4.3.8 Three-way ANOVA analysis of the average normalized score for Sentiment Analysis Model, Feature Extraction Method and Data Augmentation technique

Table 4.13 shows the results of three-way ANOVA of the average normalized score for sentiment analysis model, feature extraction method and data augmentation technique

Table 4.13: Three-Way ANOVA of the Sentiment Analysis Models, Feature Extraction Techniques and the Data Augmentation Techniques

ANOVA - Average Normalized Performance

	Sum of Squares	df	Mean Square	F	P
Model	4.03673	9	0.44853	48.83906	< .001
Feature Extraction	0.10878	1	0.10878	11.84497	0.001
Data Augmentation	0.20503	1	0.20503	22.32544	< .001
Model ✻ Feature Extraction	0.21528	9	0.02392	2.60462	0.018
Model ✻ Data Augmentation	0.18168	9	0.02019	2.19810	0.043
Feature Extraction ✻ Data Augmentation	3.12e-5	1	3.12e-5	0.00340	0.954
Model ✻ Feature Extraction ✻ Data Augmentation	0.00778	9	8.65e-4	0.09414	1.000
Residuals	0.36735	40	0.00918		

The *p*-value of less than 0.001 exhibits significant differences in the sentiment analysis models, feature extraction techniques as well as the data augmentation techniques, all at their individual levels. Moreover, there are significant interactions between model and data augmentation techniques evidenced by the *p*-value of 0.043as well as between models and feature extraction techniques evidenced by the *p*-value of 0.018.

Further post-hoc analyses demonstrate that the significant interactions between the models and the feature extraction techniques only apply to the traditional machine learning models only. These interactions do not apply to the deep learning models. On the other hand, further post-hoc analysis demonstrates that the significant interactions between the sentiment analysis models and the data augmentation

techniques only apply to the deep learning models. These interactions do not apply to the traditional machine learning models.

4.3.9 Bar Graph of the Total Average Normalized Scores of the Different Feature Extraction Methods

Figure 4.1 shows normalized performance score of the two feature extraction techniques used with the sentiment analysis models for handling the polarity shift.

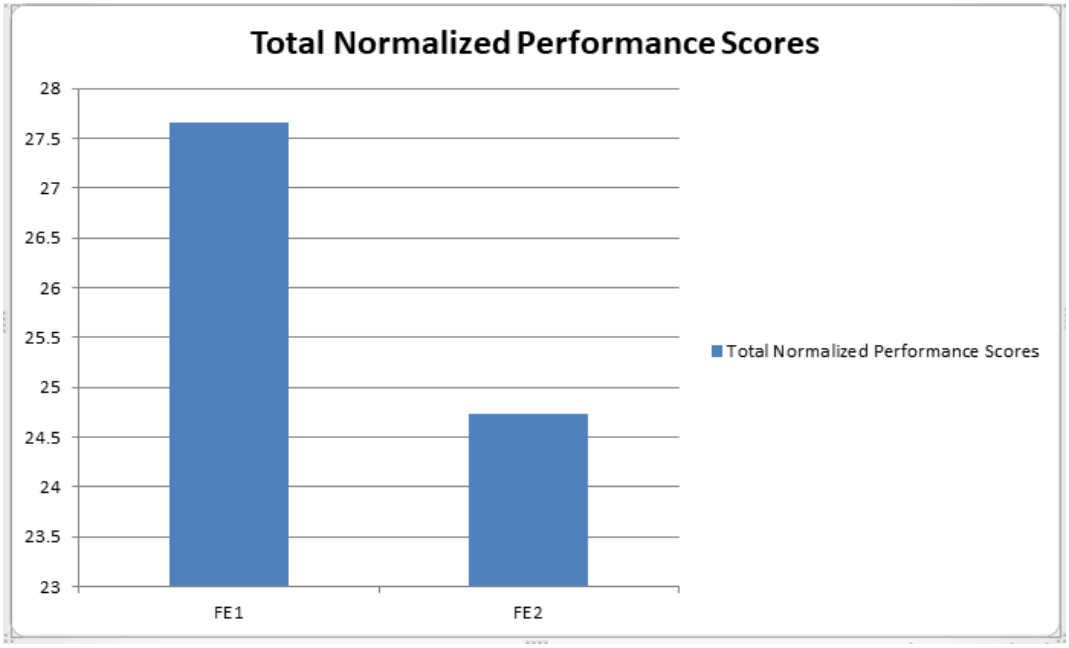


Figure 4.1: Bar graph of the total normalized average scores for feature extraction

In the figure 4.1, the acronyms for the different categories of feature extraction techniques are as follows: FE1 (FastText) and FE2 (Word2Vec). This bar graph demonstrates that the FastText feature extraction technique outperforms the word2vec feature extraction technique in the performance of the traditional models.

4.3.10 Bar Graph of the Total Average Normalized Scores of the Different Feature Selection Methods

Figure 4.2 shows the comparison of the two feature selection techniques used in the experiments. A bar graph was used to compare the total average normalized scores across different feature selection techniques used in sentiment analysis models. It is clear that feature selection plays a crucial role model's performance therefore selection of an effective technique is crucial.

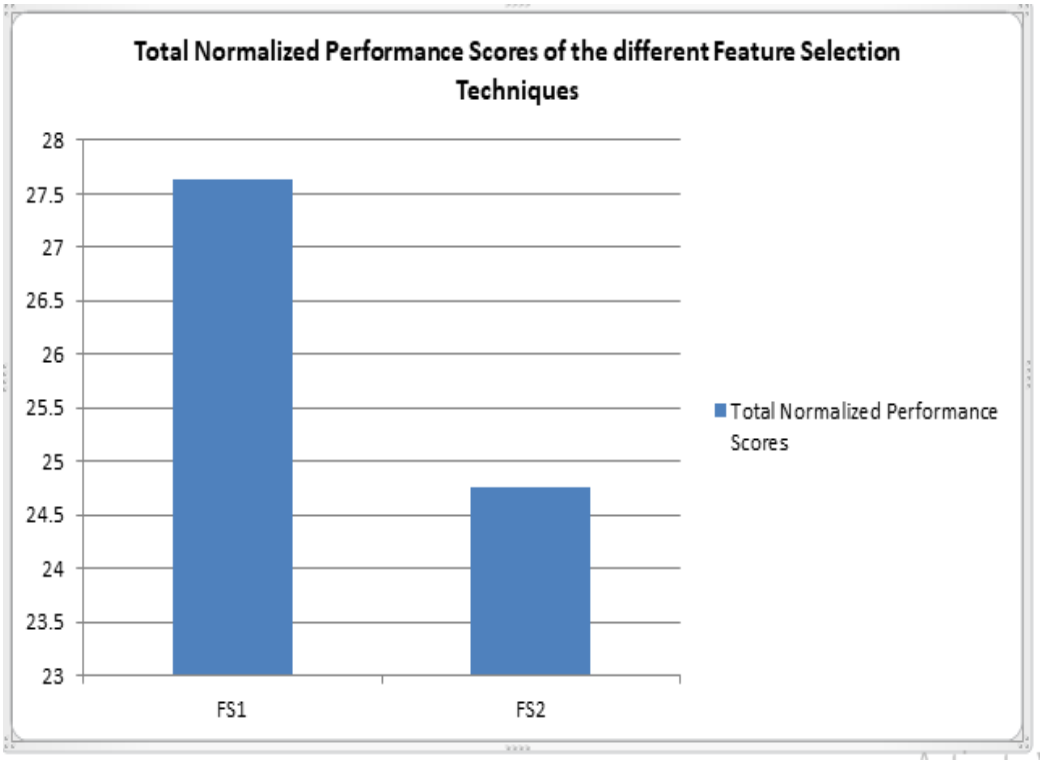


Figure 4.2: Bar graph of the total normalized average scores for the different feature selection techniques

In the above bar graph, the acronyms for the different categories of feature selection techniques are as follows: FS1 (recursive feature selection) and FS2 (Term Frequency- Inverse Document Frequency). This bar graph demonstrates that the recursive feature selection (RFE) technique surpasses the performance of the TF-IDF (Term Frequency- Inverse Document Frequency) in the performance of the traditional machine learning based sentiment analysis models. Feature selection technique like recursive feature elimination show higher normalized scores, indicating their effectiveness in selecting relevant features for sentiment

classification. The differences in scores is clear indication of how certain techniques contribute more significantly to improving model's classification accuracy.

4.3.11 Bar Graph of the Total Average Normalized Scores of the Different Data Augmentation Techniques

Figure 4.3 shows a bar graph of the comparison of the total average normalized scores across different data augmentation techniques used in the empirical analysis of the different sentiment analysis models.

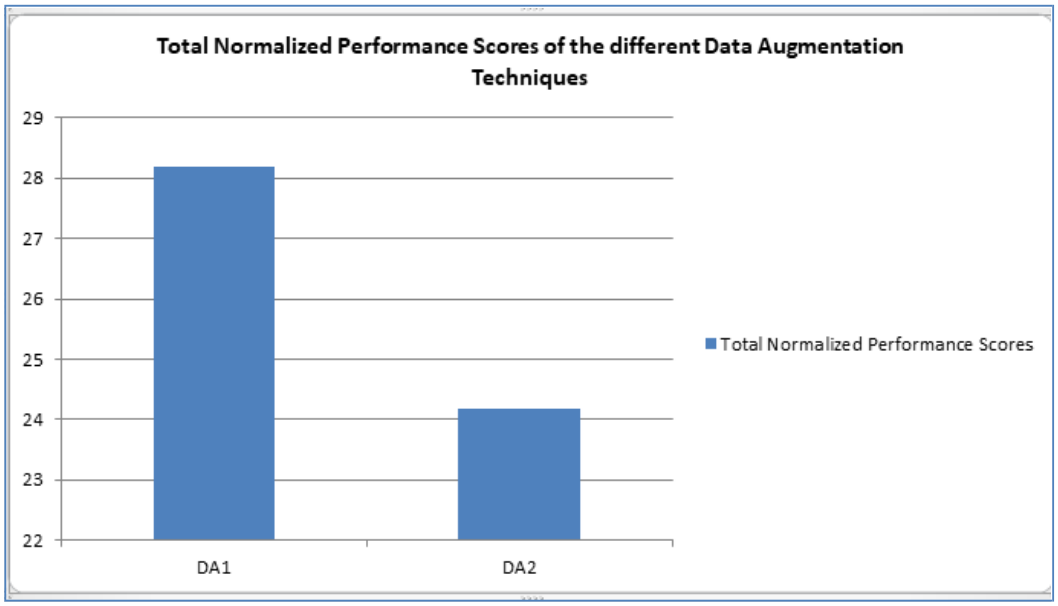


Figure 4.3: Bar graph of the total normalized average scores for the different data augmentation techniques

In the figure 4.3 the acronyms for the different categories of data augmentation techniques are as follows: DA1 (Back translation) and DA2 (synonym replacement). This bar graph demonstrates that the Back translation data augmentation technique outperforms the synonym replacement technique in the performance of the deep learning based sentiment analysis models. The graph presents the performance variations among data augmentation by enhancing the model's performance. Therefore, this increases the diversity of training data, and hence helps in the mitigation of class imbalance and overfitting issues with the sentiment classification models.

4.3.12 Bar Graph of the Total Average Normalized Scores of the Different Categories of Sentiment Analysis Models

Figure 4.4 shows the normalized performance score across different categories of sentiment analysis models. The main is to analyze the performance variations among different model types in sentiment classification tasks.

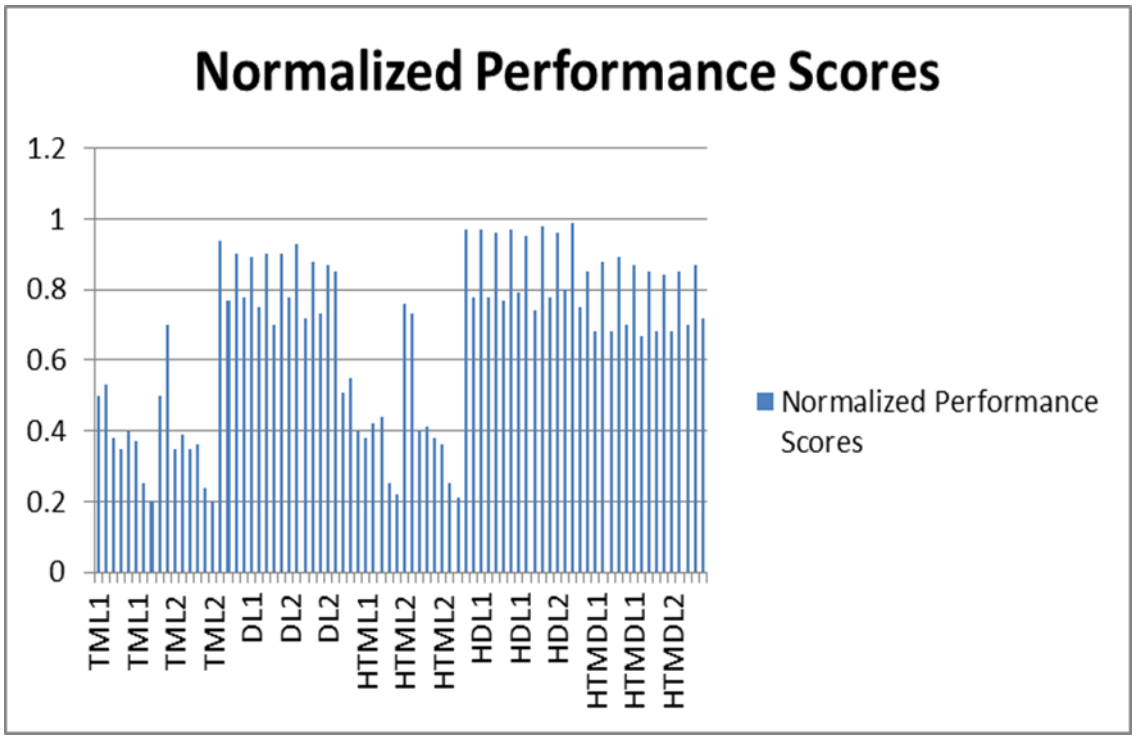


Figure 4.4: Bar graph of normalized performance scores of sentiment analysis models

In the above bar graph, the acronyms for the different categories of models refers to the follows: TML1 (first traditional machine learning model), TML2 (second traditional machine learning model), DL1 (first Deep learning model), DL2 (second Deep learning model), HTML1 (first hybrid made up of traditional machine learning models exclusively), HTML2 (second hybrid made up of traditional machine learning models exclusively), HDL1 (first hybrid made up of deep learning models exclusively), HDL2 (second hybrid made up of deep learning models exclusively), HTMDL1 (first hybrid made up of both the traditional and deep learning models) and lastly, HTMDL2 (second hybrid made up of both the traditional and deep learning models). This graph demonstrates that deep learning models and their respective hybrids exhibit the highest total average normalized scores compared to

the traditional machine learning models and their hybrids. This is due to their superior ability to capture complex linguistic patterns. The BERT-based model often outperforms others due to their contextual understanding of text.

4.4 Discussion

This empirical research delves into the experimentation and performance evaluation of diverse sentiment analysis models, broadly categorized into five groups: traditional machine learning models, deep learning models, hybrid models integrating both machine learning and deep learning, hybrid models predominantly utilizing deep learning methodologies, and hybrid models exclusively employing machine learning approaches. The empirical analysis unravels the impact of various feature extraction, feature selection, and data augmentation techniques on the models' efficacy in handling polarity shifts. Through rigorous experimentation, advanced ANOVA statistical analyses are employed using the R software to extract key findings and insights.

4.4.1 Significant Differences between Traditional ML and Deep Learning Models

The ANOVA analysis reveals notable performance distinctions between traditional machine learning (ML) models and deep learning (DL) models. It is also notable that the architectures and methodologies employed in deep learning models are particularly effective in handling polarity shifts compared to traditional machine learning approaches. Deep learning models inherently capture more complex patterns in sentiment analysis tasks, which are crucial for handling nuanced sentiment flux. These observations are in line with the ones made by Pathak et al. [178]. Research shows that deep learning approaches benefit from unsupervised representation learning inorder to obtain unprecedented general language knowledge, using a large amount of unlabelled data [179]. These empirical observations offer foundational insights that pave the way for the future development of an advanced sentiment analysis model. A transformer-based BERT-based deep learning model that strategically configures hyperparameter for effective feature extraction within the deep network was incorporated.

4.4.2 Interactions between Sentiment Analysis Models and Feature Extraction Techniques

The results of the experiments show there is significant interactions between traditional ML models and feature extraction techniques. Notably, interactions between traditional machine learning models and feature extraction techniques underscore the pivotal role played by these techniques in shaping model performance. The analyses reveal that most feature extraction in deep learning models occurs within the deep network itself, minimizing the influence of external feature extraction techniques. Therefore there is need to optimize the feature extraction process while using traditional machine learning approaches. This observations in line with the one made by Symeonidis et al. [180]. Deep learning models are capable of automatically learning hierarchical representations of text, eliminating the need for explicit feature engineering. This finding underscores the self-sufficiency of deep learning models in extracting meaningful features from raw data. These observations are in line with the ones made by Leng and Jiang [181].

Additionally, the results show a superior performance of Fast Text feature extraction technique as opposed to Word2Vec. This observations are in line with the one made by Hasan et al. [182]. This is attributed to the fact that FastText considers subword information (character n-grams) when generating word embeddings. This means that even for words that are not present in the training corpus or are out-of-vocabulary (OOV), FastText can still create meaningful embeddings based on their constituent character sequences. In sentiment analysis, words affected by implicit negation usually have variations due to prefixes, suffixes, or internal character changes. The inclusion of subword information enhances the model's generalization capabilities. These qualities contribute to the generation of more robust features for sentiment analysis models dealing with polarity shift and implicit negation in datasets like Amazon mobile reviews.

4.4.3 Feature Selection Techniques Impact on Sentiment Analysis

The analysis results show there is an interaction between traditional ML models and feature selection techniques. Traditional ML models benefit from selecting relevant features to improve their discriminative power in sentiment analysis tasks, indicating the importance of feature selection strategies in optimizing model performance.

These observations are in line with the ones made by Cai et al. [183]. The finding suggests that deep learning models inherently learn to focus on informative features during training, reducing the need for explicit feature selection techniques. These observations are in line with the ones made by Zou et al. [39].

The observation of superior performance of the wrapper-based recursive feature selection (RFE) technique as opposed to the TF-IDF is in line with the observation made by Onan [184]. This is attributed to the RFE's customization, model performance optimization, interpretability, and alignment with domain-specific nuances, making it well-suited for sentiment analysis tasks focused on handling polarity shifts in such datasets. Amazon reviews dataset contains nuanced language, domain-specific terms, and implicit sentiment expressions influenced by negation.

4.4.4 Data Augmentation Techniques and its Impact in Sentiment Analysis Model

The results show absence of significant interactions between traditional ML models and data augmentation techniques. This indicates that data augmentation methods may not significantly impact the performance of traditional ML models in handling polarity. This suggests that traditional ML models may not benefit as much from data augmentation techniques compared to deep learning models. These observations are in line with the ones made by Sakai et al. [185]. On the hand, there is a significant interaction between deep learning models and data augmentation techniques. This highlight the importance of data augmentation in enhancing the performance of deep learning models for sentiment analysis. Specifically, the superior performance of the back translation method compared to synonym replacement suggests that generating diverse and semantically meaningful augmented data through back translation contributes significantly to improving the robustness of deep learning models in capturing polarity shifts. These observations are in line with the ones made by Lee [186].

The observation of the higher performance of the Back translation as compared to the Synonym replacement is in line with the one made by Beddiar et al. [187].This is attributed to the fact that back translation introduces significant semantic variations by translating sentences into another language and then back to the original. This variation helps deep learning models learn diverse linguistic patterns and capture

implicit nuances in sentiment. Deep learning models, particularly those based on transformers like BERT, rely heavily on contextual information. By leveraging back translation, the sentiment analysis model can learn from a more diverse set of examples and generalize well to unseen data scenarios within the Amazon mobile reviews domain.

Developing context-aware back translation approaches that consider the context of the original text and try to preserve idiomatic expressions or domain-specific language can help mitigate this challenge. Poor translations can introduce noise or incorrect sentiment cues into the augmented data, affecting model performance. Leveraging multiple translation models and incorporating quality checks can also enhance the reliability of the augmentation process. Overall, while back translation is a powerful augmentation technique, addressing these challenges through advancements in translation models, context-aware approaches, quality checks, and domain adaptation strategies can further enhance its effectiveness in handling polarity shifts and improving sentiment analysis model performance on diverse datasets like Amazon mobile reviews. Among the challenges mentioned, one major challenge in handling datasets used by the sentiment analysis models used to handle polarity shift due to implicit negation through back translation is preserving specific context and idiomatic expressions.

However, back translation may not always capture these aspects effectively, leading to a loss of context and potentially affecting model performance. Improvement through perturbations can help address this challenge by introducing controlled variations in the back-translated data while preserving important contextual and idiomatic elements. Perturbations involve making small modifications or additions to the data to create diverse yet contextually relevant examples. The hybrid model combines perturbations with the back translation data augmentation strategy to further enrich the training data with diverse linguistic patterns and sentiment expressions. Combining Seq2seq models with perturbation techniques during back translation helps in creating more diverse and contextually relevant augmented datasets for training NLP models, enhancing their performance in tasks like sentiment analysis on datasets such as Amazon reviews.

4.5 Key Findings from Evaluation Analysis

The key findings from the evaluation analysis and their implications in the research work are discussed in this section. These finding are as a result of the experiment data analysis.

There are no statistically significant interactions between deep learning models and feature extraction techniques. The choice of feature extraction techniques does not significantly impact the performance of deep learning models, allowing machine learning developers to focus on other optimizations.

There are no statistically significant interactions between deep learning models and feature selection techniques. Feature selection does not significantly influence deep learning model performance, so researchers can prioritize other hyperparameter tuning and model architecture improvements. There are statistically significant interactions between deep learning models and data augmentation techniques. Data augmentation significantly impacts model performance, highlighting the need for careful selection of augmentation strategies, particularly for handling polarity shifts.

There are statistically significant interactions between traditional machine learning models and feature extraction techniques. Feature extraction methods play a crucial role in improving the performance of traditional machine learning models and should be carefully selected, in traditional machine learning algorithms, based on dataset characteristics. There are statistically significant interactions between traditional machine learning models and feature selection techniques. Proper feature selection techniques are essential for optimizing traditional machine learning models, as they significantly affect model performance.

Data augmentation does not have a significant impact on traditional machine learning models, and efforts should be focused on feature selection and model tuning instead. RFE should be prioritized over TF-IDF in feature selection tasks, particularly in scenarios requiring model interpretability and optimal feature subset selection. While FastText should be preferred for feature extraction in sentiment analysis as it captures subword information, improving generalization to unseen words. On the other hand, back translation should be the preferred data augmentation

technique to enhance model robustness, as it preserves semantic meaning while introducing diversity.

Deep learning should be prioritized over traditional machine learning for sentiment analysis tasks, particularly when dealing with complex linguistic structures like implicit negation. Across all the experiments, hybrid models performed better than standalone algorithms. Hybrid models should be adopted for complex sentiment analysis like the one for implicit negation, as they leverage the strengths of multiple architectures to improve accuracy and robustness. Machine learning developers for sentiment analysis should aim at combining multiple deep learning models results in superior performance compared to hybridizing deep learning with traditional machine learning models.

4.6 Summary

Chapter four presents a rigorous empirical investigation into the effectiveness of sentiment analysis models in detecting polarity shifts due to implicit negation. The findings demonstrate that the existing models struggle with nuanced sentiment shifts due to implicit negation, necessitating the development of more advanced hybrid approaches. The pilot study establishes foundational insights, guiding the selection of models applied in the empirical analysis. The empirical analysis process culminates to a set of key findings and a respective practical implication for each and every key finding. One of the practical implications is the use of contextual back translation and Seq2Seq perturbations in the development of a state of the art hybrid deep learning model for detecting polarity shift due to implicit negation. Overall, this chapter establishes a strong empirical foundation for the proposed hybrid deep learning model. The statistical and visual analyses provide compelling evidence of its effectiveness, setting the stage for further discussions in the subsequent chapter five regarding its deployment in real-world sentiment analysis applications. Future research directions, such as optimizing computational efficiency through knowledge distillation, pruning, and quantization, are suggested, in chapter six, as potential avenues for enhancing model's performance while reducing computational resource requirements

CHAPTER FIVE

HYBRID DEEP LEARNING MODEL DEVELOPMENT

5.1 Introduction

The theoretical analysis in chapter two show that the existing sentiment analysis models often struggle to accurately discern the sentiment polarity of texts containing subtle negations, such as "not bad" or "hardly impressive," leading to misclassifications. This challenge is pronounced in the user-generated content, where implicit negations frequently appear and shape consumer sentiment in complex ways. Consequently, the empirical analysis in chapter four, based on ANOVA, demonstrates that the efficacy of deep learning based sentiment analysis models is significantly influenced by the choice of the data augmentation techniques. To address this gap, this chapter presents the design and development of a novel hybrid deep learning model. The model leverages advanced data augmentation techniques, contextual embeddings, and attention mechanisms to enhance sentiment classification accuracy in the presence of implicit negation. Section 5.2 introduces a multi-methodological model development methodology. Section 5.3 details the model architecture and development process, encompassing data preprocessing, augmentation strategies, feature extraction and the integration of attention mechanisms.

5.2 Model Development Methodology

This stage involves the design and development of the model. This involves the understanding of the conceptual framework, data collection and preprocessing. During the data preprocessing, the features are standardized in order to ensure that they have similar scales, as this improves the training. The data pre-processing also entails handling any missing values appropriately, either by imputing them or encoding them as a separate category. The model design and development focusses on the setting up of detailed specifications for the model including its architecture and its components. The model training and hyperparameter are based on Cross-Entropy Loss and Adam optimizer respectively. The Adam is used as the adaptive learning rate strategy to dynamically adjust the learning rates based on the gradients. This ensures model’s performance is validated with this learning rate on a validation set, monitoring both the training and the validation loss curves. The learning rate

provides the best trade-off between the convergence speed and the performance. Figure 5.1 shows the model development methodology that is adopted in the development of the hybrid deep learning model for detecting implicit negations.

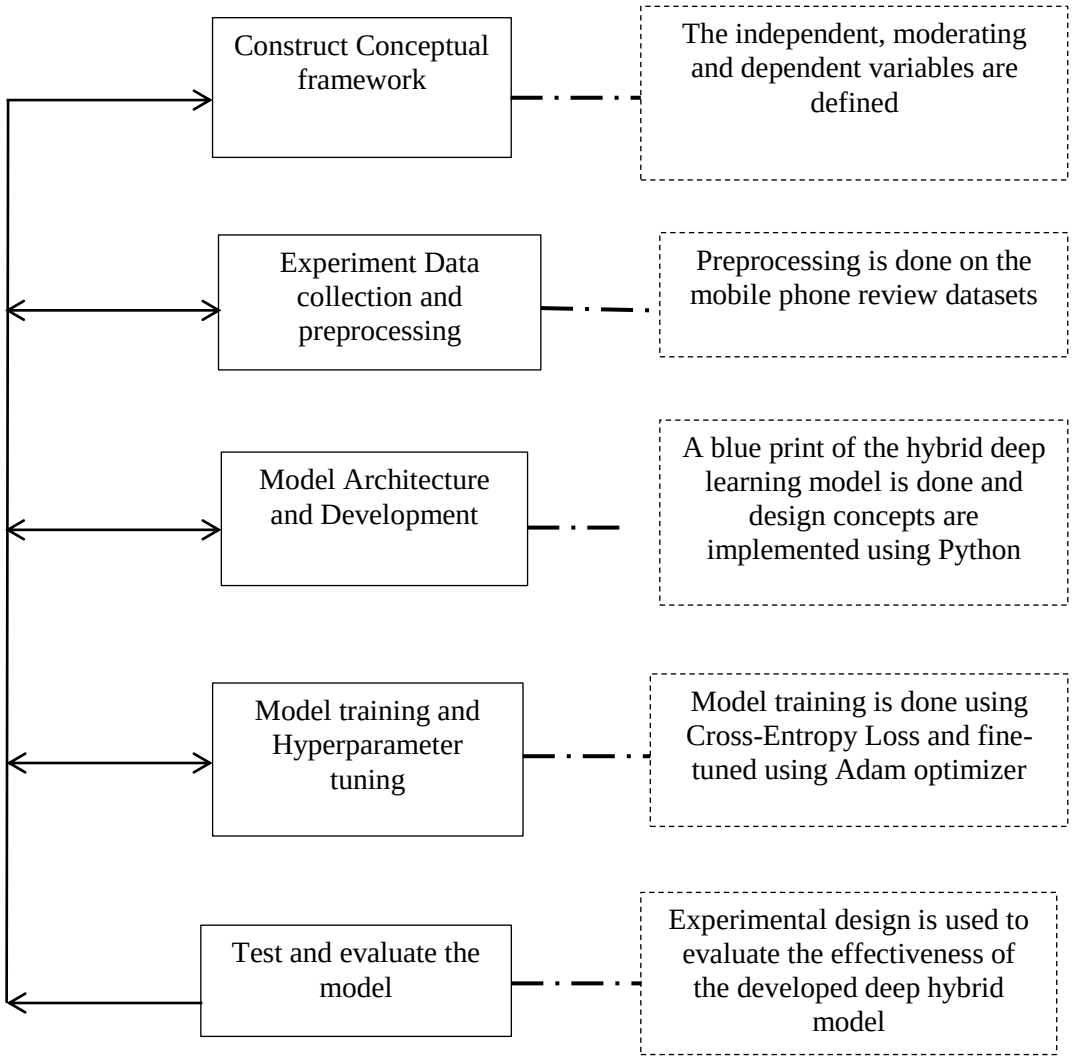


Figure 5.1: Model Development Methodology

5.3 Model Architecture Design

Based on the data analysis results of the experimental data from the main experiments, the newly developed hybrid deep learning model is designed and implemented using a multi-methodological approach. The hybrid model integrates Seq2Seq perturbations and contextual back translation to enrich training data diversity while preserving semantic integrity. Furthermore, it combines FastText and BERT embeddings to capture both subword-level and deep contextual representations, thereby enhancing the model's ability to detect nuanced polarity

shifts. The core architecture is centred on a transformer-based and LSTM framework with an attention mechanism that selectively focuses on sentiment-critical tokens, and ensuring that implicit negations are effectively identified and processed. Additionally, the model employs a sophisticated training strategy, incorporating adaptive learning rate scheduling and dropout regularization to optimize convergence and generalization performance. The transformer-based BERT and LSTM hybrid deep learning model strategically configures hyperparameters for effective feature extraction within the deep network. The experimental setup follows a structured approach, incorporating:

i. Data preprocessing- this involved stop-word removal using the natural language toolkit library, tokenization and lemmatization to normalize textual data and finally noise reduction is done via regex-based filtering. This ensured that the features are standardized to similar scales inorder to improve the training process.

ii. Feature Engineering- FastText and BERT embeddings are used for contextual understanding while TF-IDF is used to capture word co-occurrence.

iii. Model Training & Hyperparameter Tuning- the batch size optimization included (16, 32) via grid search. The learning rate was set to 2e-5 with dynamic scheduling using Adam optimizer. Lastly the epochs are set to 100 but the early stopping is used to prevent overfitting.

iv. Evaluation Metrics & Statistical Validation- In this stage of evaluation, the ANOVA and Kruskal-Wallis H tests assessed the statistical significance. Attention Heatmaps and whisker plots are used for visualizations. SHAP visualizes the developed model's interpretability.

This structured approach ensures methodological rigor and reproducibility hence improving sentiment classification robustness.

5.4 Model Development

In this study, we focused on developing a hybrid deep learning models that can effectively detect and classify implicit negation using the amazon mobile phone reviews datasets. The model development process involved three key stages: data preprocessing, model architecture and model training. Particular emphasis was placed on ensuring the model could generalize well to unseen data, avoid overfitting,

and remain computationally efficient. The transformer-based model architectures with attention mechanism and feedforward layers are explored in this development. Additionally, we integrated techniques such as contextual back translation augmented with seq2seq perturbation for data augmentation process to improve dataset during training. This subsequent subsection outlines the data preprocessing stage, the detailed architectural design, and the training procedure employed to build the final model.

5.4.1 Data Preprocessing of the Proposed Hybrid Model

The hybrid deep learning model for handling polarity shifts due to implicit negation in sentiment analysis was a comprehensive and innovative approach, specifically tailored for nuanced challenges in analysing mobile phone brand reviews. By integrating advanced data augmentation, feature handling, and architectural strategies, this model addresses the persistent challenge of implicit negations like "not bad" or "hardly impressive," . The model's design reflects a systematic workflow that ensures robustness, adaptability, and high performance. Figure 5.2 shows the data preprocessing steps undertaking in the development of the hybrid deep learning model.

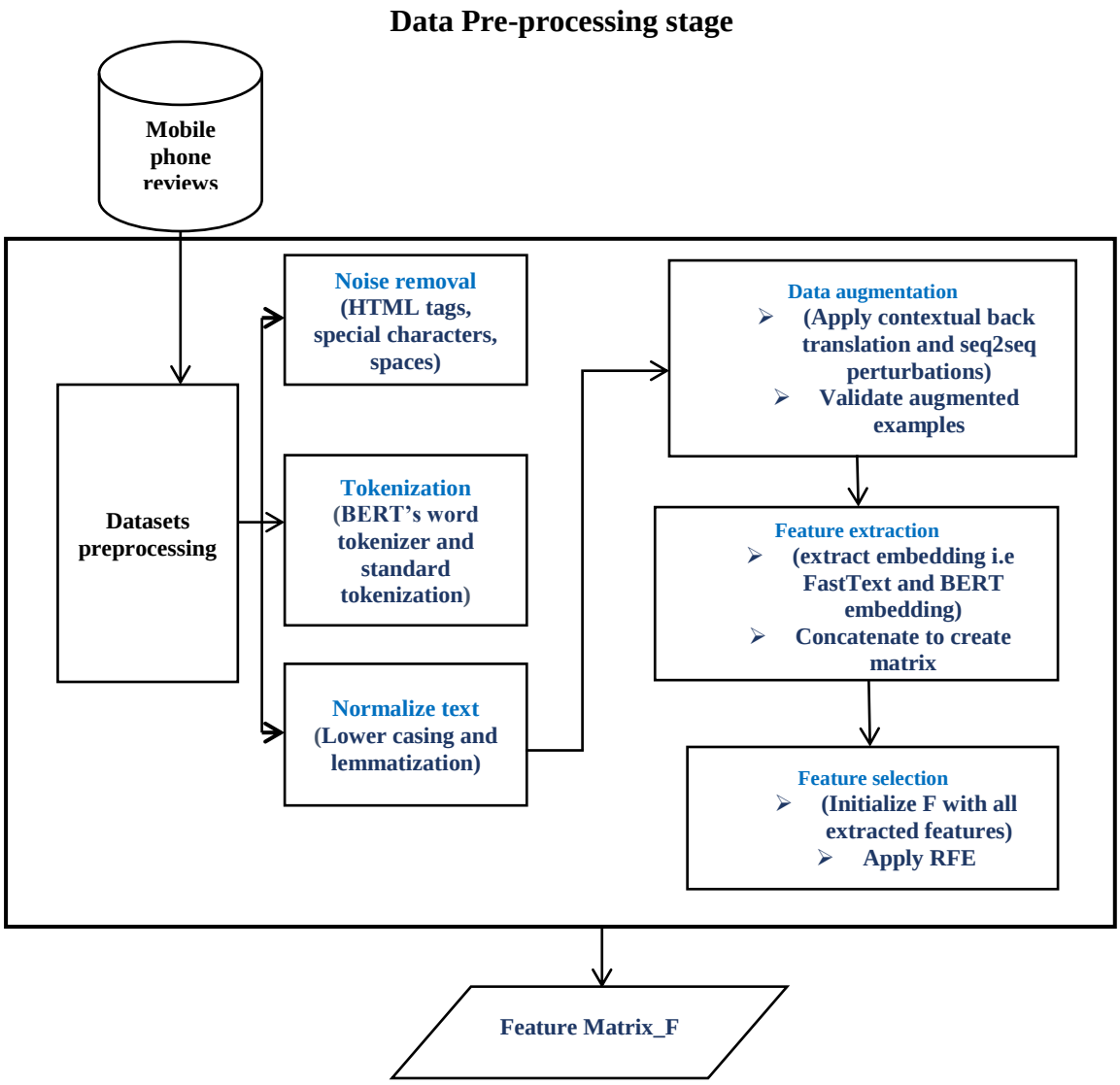


Figure 5.2: Data Preprocessing of the Hybrid Model

The data preprocessing stage involves a number steps that aimed to ensure a standardized dataset, enabling downstream processes to operate efficiently. This study utilized the amazon mobile phone datasets that contained 67,987 records with annotations of values representing negative, positive and neutral polarities. Negative sentiment polarity was annotated with values ranging between 1 and 2. The positive sentiment polarity was annotated with values ranging between 4 and 5. Lastly the neutral sentiment polarity was annotated with value 3. Firstly, the noise such as HTML tags, special characters, and redundant spaces are removed. The preprocessing phase also includes tokenization and normalization to handle variations in linguistic patterns, crucial for accurately capturing the context of negations. Secondly, the model implements a data augmentation strategy that

combines contextual back translation and Seq2Seq perturbations. Back translation generates diverse linguistic variations by translating text into an intermediate language and back to the original, preserving the meaning while altering syntax. This method enhances diversity but, on its own, may fail to capture certain nuances like idiomatic expressions or domain-specific terms. To address this, Seq2Seq perturbations are introduced. The data is passed through Marian MT Seq2Seq encoder for encoding and a beam search is used to decode multiple syntactic variations. These perturbations create controlled syntactic variations, ensuring semantic integrity while enriching the training data with contextually relevant and linguistically diverse examples. Quality control measures, such as automated validation or manual checks, are embedded within this process to maintain authenticity and relevance. Thirdly, feature extraction process is done, where the model leverages a hybrid approach combining FastText embeddings and BERT embeddings. FastText captures semantic relationships at the subword level, making it effective for handling spelling variations and domain-specific terms often found in reviews. In contrast, BERT embeddings provide deep contextual representations, enabling the model to understand complex linguistic structures and nuances in negation. Extracted features are refined using Recursive Feature Elimination (RFE), which iteratively selects the most relevant features by removing those with lesser predictive power. This step reduces dimensionality, enhances computational efficiency, and ensures that only critical features influencing polarity shifts are retained. (77)

5.4.2 Model Architecture of the Proposed Hybrid Deep Learning Model

The core architecture of the model integrates multiple components to comprehensively address the polarity shift problem. BERT serves as the backbone, providing robust contextual understanding, while RFE optimizes feature selection. Figure 5.3 describes the details of the hybrid model architecture.

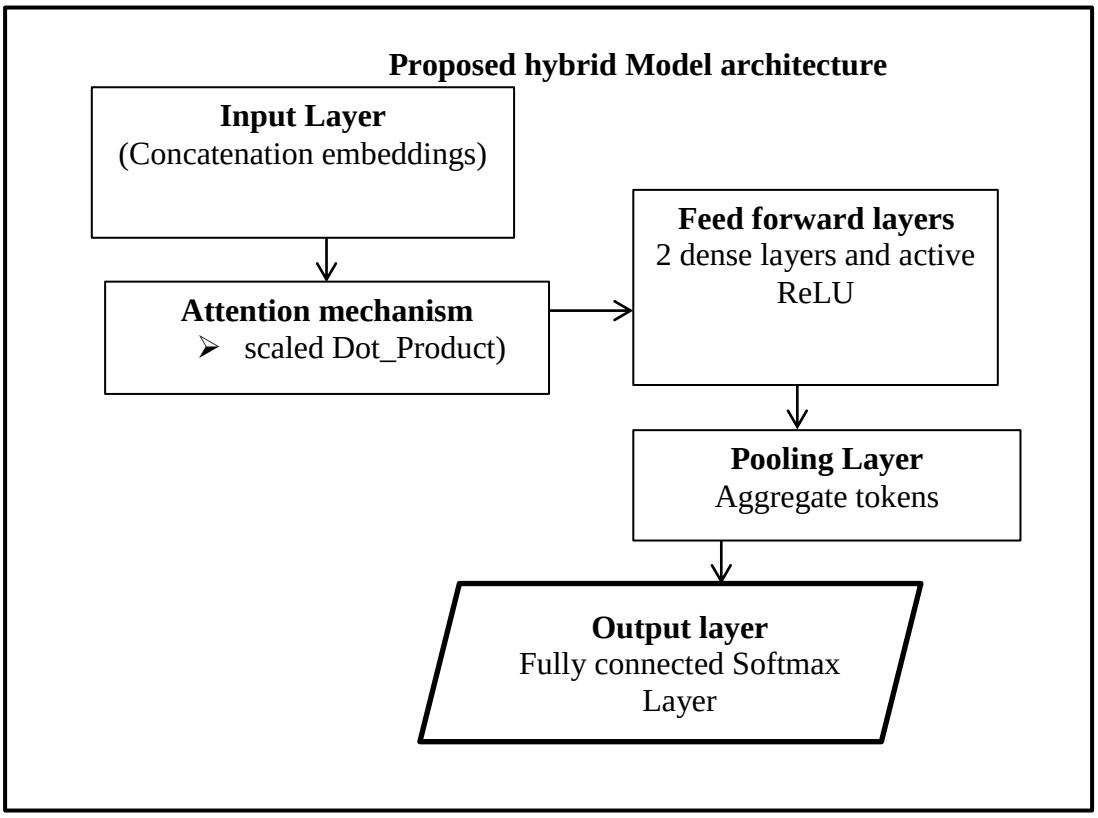


Figure 5.3 Proposed Hybrid Model Architecture

The hybrid model incorporates an attention mechanism, specifically scaled Dot-product attention to highlight critical sentiment words and negations, assigning appropriate weights to tokens based on their relevance. The model also incorporates Seq2Seq perturbations into its architecture to enhance resilience to syntactic variations. The final architecture includes feedforward layers with ReLU activation for further feature refinement and mean pooling to aggregate token embeddings, preserving key contextual information. Predictions are generated through a fully connected softmax layer, converting logits into probabilities for multi-class sentiment classification. The combination of FastText and BERT embeddings offers complementary strengths, capturing both semantic subtleties and deep contextual relationships**.** The architecture's use of attention mechanisms and RFE further enhances its capability to focus on critical sentiment-bearing features, while quality control in augmentation ensures the integrity of the training data.

These innovations collectively enable the model to handle the polarity shift problem more effectively than traditional methods, setting a new benchmark in sentiment analysis. Its real-world relevance is underscored by its application to mobile phone reviews, making it invaluable for consumer analytics and e-commerce sentiment analysis. This comprehensive design ensures the model is robust, adaptable, and capable of addressing nuanced challenges in sentiment analysis, particularly implicit negations. It serves as a significant advancement in the field, offering a framework that can be extended to other domains requiring fine-grained sentiment understanding.

5.4.3 Model Training

In order to improve on the training performance, learning rate scheduling and early stopping techniques are used. Training is conducted with Cross-Entropy Loss, which is ideal for multi-class classification tasks, ensuring effective differentiation between sentiment classes. The optimization theory underpins the training process which aimed at optimizing loss function inorder to minimize errors. The model uses the Adam optimizer, fine-tuned with a learning rate of 2e-5 for efficient convergence**.** Learning rate scheduling dynamically adjusts the rate during training to prevent overshooting or stagnation. Training is performed over 100 epochs, with a batch size of 32, balancing computational efficiency and gradient stability. A dropout rate of 0.3 mitigates overfitting by deactivating neurons randomly during training, enhancing generalization.

Figure 5.3 shows the model framework for the training and testing processes. The key processes in the model training and testing are highlighted in the framework.

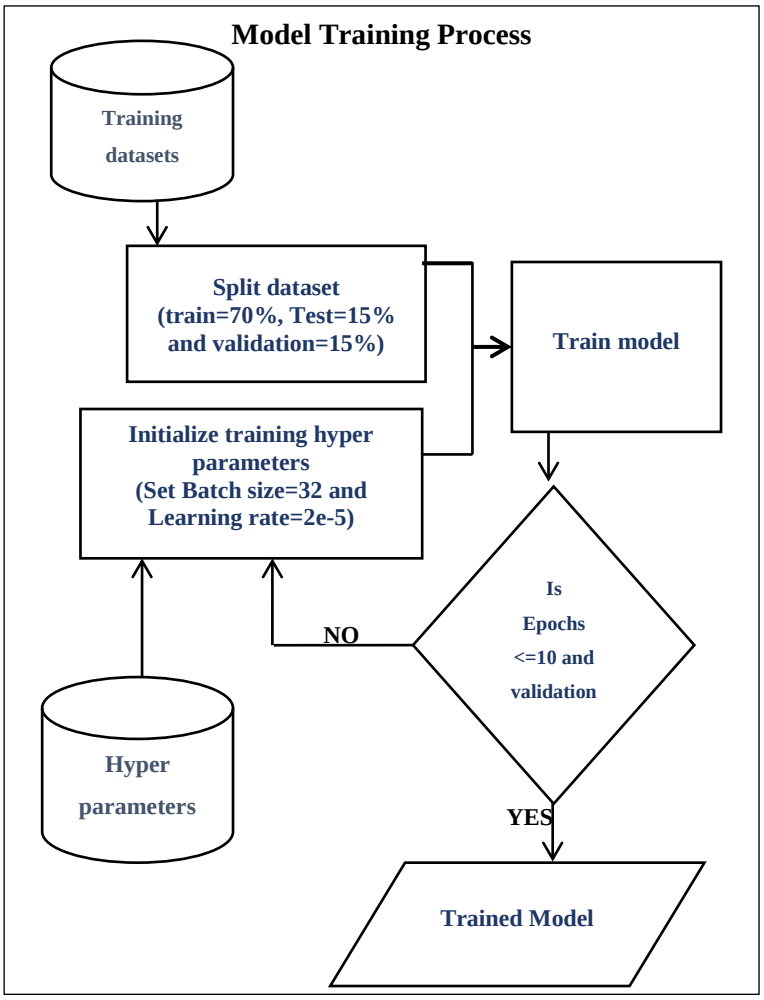


Figure 5.4: Model Training, and Testing Framework.

During the training process, the initial hyperparameter is set to a dropout layer with a rate of 0.3 after key hidden layers. In this, the co-dependence among neurons is reduced by randomly deactivating 30% of neurons during training to counter overfitting. The input data was divided into batches of 32 that allowed for faster computations and more stable gradients estimates. The model performed a forward pass for each batch, computing predictions based on current weights. Dropout was further was applied during this phase, ensuring certain neurons were randomly turned off, introducing regularization. At this point, the predicted outputs were compared to the ground truth labels using a cross-entropy loss function. Adam optimizer updated the model parameters in order to adjust the weights based on the gradients and learning rate. This process was repeated for all batches in the epoch.

The model's performance was evaluated on a validation set to monitor training progress and detect any cases of overfitting. In addition to dropout, other regularization techniques such as early stopping to monitor validation loss were employed to prevent unnecessary training and model degradation. Training was continued for up to 100 epochs and the loss and validation curves of the hybrid models are shown in figures 5.5 and 5.6

The model loss curves for the proposed hybrid deep learning model, as shown in the graph below, demonstrate a typical training process where both training and validation losses start high and gradually decrease over time before plateauing. This behavior provides valuable insights into the model's learning process and its effectiveness in handling the task. At the beginning of the training, both the training and validation losses start at relatively high values. This is because the model begins to learn from the data, and the initial weights are set randomly, resulting in poor predictions.

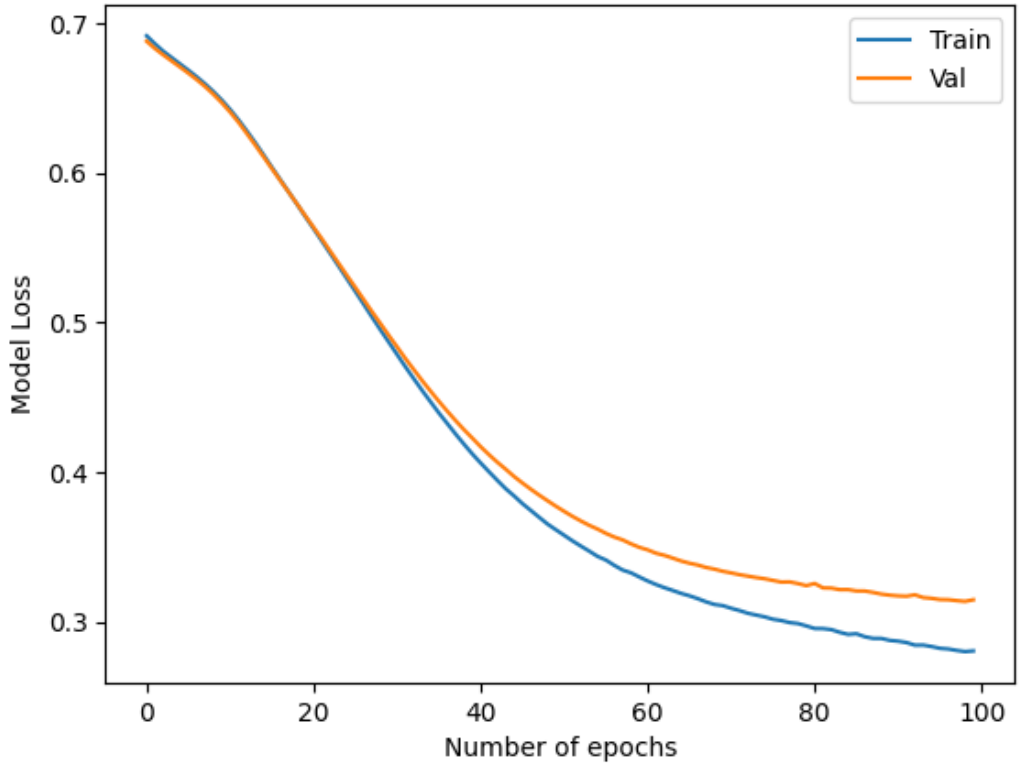


Figure 5.5: Loss of the Hybrid Deep Learning Model for Detecting Polarity Shift

During this early phase, the model explores the data space without a clear understanding of the patterns, leading to higher losses. As the hybrid model continues to train, both the training and validation losses steadily decrease. This decline indicates that the model is improving its ability to predict sentiments, with the weights being adjusted to minimize the error between predicted and actual values of the sentiments due to implicit negation. The training loss curve drops at a faster rate compared to the validation loss curve, which suggests that the model is fitting to the training data more quickly than to the validation data. This is typical as the model initially learns from the examples it has seen most frequently during training. Eventually, after around 60 epochs, both the training and validation loss curves begin to plateau. This suggests that the model has reached its optimal learning point, where further training upto 100 results in minimal improvements.

The plateau indicates that the model is no longer significantly improving, and the learning process has effectively converged. This is an important milestone, as it shows that the model has learned as much as it can from the available data. The close tracking of the training and validation loss curves as they plateau is a strong indication that the model is generalizing well to unseen data. If there had been a large gap between the training and validation losses, it would have signified overfitting, where the model performs well on the training data but poorly on new, unseen data. However, in this case, the losses level off in tandem, suggesting that the hybrid deep learning model is not overfitting and is maintaining its performance across both the training and validation datasets. Overall, the gradual decline followed by the plateau in the loss curves reflects that the hybrid deep learning model has learned effectively. The stable training process indicates that the model is not only fitting the training data but is also generalizing well, which is essential for accurately addressing the polarity shift challenge due to implicit negation in sentiment analysis, as outlined in this research.

The graph in figure 5.6 below depicts training and validation average performance score over epochs. It shows a typical and desirable learning process for the newly developed hybrid deep learning model for detecting polarity shift due to implicit negation. Both the training and validation average performance score curves start at low values and gradually increase as the model progresses through each epoch.

This gradual increase signifies that the model is effectively learning from the data and improving its ability to make accurate predictions.

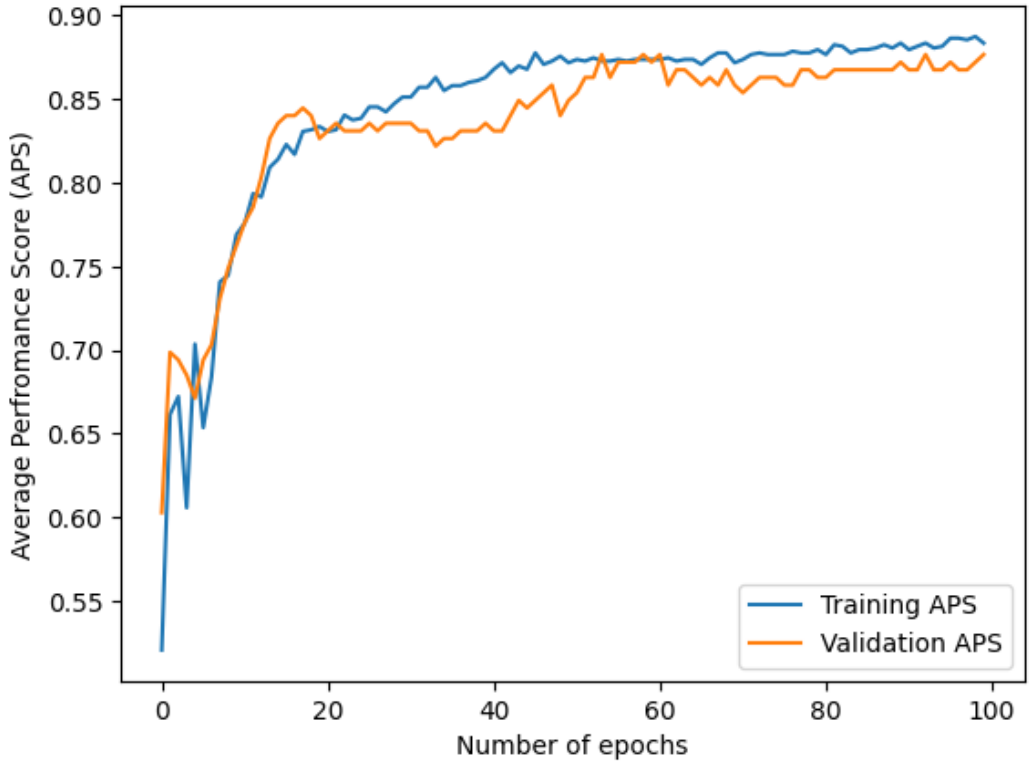


Figure 5.6: APS of the Hybrid Deep Learning Model for detecting polarity shift

During this phase, the model adjusts its weights and optimizes the parameters, leading to a higher level of understanding of the underlying patterns in the data. As the training continues, both the training and validation average performance score curves show an upward trajectory, albeit a slight decline between 5 and 10 epochs. The increase in average performance score indicates that the model is learning to generalize better, correctly classifying the sentiment and making accurate sentiment predictions.

The validation average performance score curve, while increasing at a slightly slower pace than the training average performance score curve, still mirrors the general trend, indicating that the model is not overfitting to the training data and is maintaining its ability to generalize on unseen data. Eventually, both curves begin to plateau, reaching high average performance score values. This plateau suggests that the model has converged to a point where further training does not result in significant improvements in average performance score.

The model has successfully learned the essential features of the data and is now performing optimally in terms of average performance score. The fact that both curves plateau at a high average performance score level indicates that the model is well-trained and has effectively captured the underlying patterns, making it capable of accurately predicting sentiment in the data. Overall, the gradual increase in average performance score followed by the plateau at high levels of performance reflects a well-functioning and generalizable model.

5.4.4 Model's Algorithm

The algorithm for the proposed model is represented as follows:

BEGIN

Step 1: Data Preprocessing

Input: Raw dataset D (text reviews)

Output: Preprocessed dataset D_clean

1. For each text T in D:
   a. Remove noise (e.g., HTML tags, special characters, redundant spaces).
   b. Tokenize text using:
      i. BERT's WordPiece tokenizer for contextual embedding compatibility.
      ii. Standard tokenization for FastText embedding preparation.
   c. Normalize text by lowercasing and lemmatizing words.

Step 2: Data Augmentation

Input: Preprocessed dataset D_clean

Output: Augmented dataset D_augmented

1. For each text T in D_clean:
   a. Apply contextual back translation:
      i. Translate T to an intermediate language (e.g., French) and back to the original.
      ii. Retain only examples preserving sentiment and semantic integrity.
   b. Apply Seq2Seq perturbations:
      i. Pass T through a Seq2Seq encoder (e.g., MarianMT).
      ii. Decode multiple syntactic variations using beam search.
   c. Validate augmented examples to ensure:
      i. Semantic consistency.
      ii. Retention of domain-specific terms, idiomatic expressions, and cultural relevance.

2. Combine original texts and validated augmented texts into D_augmented.

Step 3: Feature Extraction

Input: Augmented dataset D_augmented

Output: Feature matrix F

1. Extract embeddings:

a. Generate FastText embeddings for subword-level semantic features (300 dimensions).

b. Generate BERT embeddings for contextual features (768 dimensions).

2. Concatenate FastText and BERT embeddings to create the feature matrix F.

Step 4: Feature Selection

Input: Feature matrix F

Output: Refined feature matrix F_refined

1. Initialize F with all extracted features.

2. Apply Recursive Feature Elimination (RFE):

a. Train a classifier to rank feature importance.

b. Iteratively remove the least important features.

c. Stop when an optimal subset F_refined is reached.

Step 5: Hybrid Model Architecture

Input: Refined feature matrix F_refined

Output: Sentiment predictions

1. Input Layer:

a. Accept concatenated embeddings from F_refined.

2. Attention Mechanism:

a. Apply Scaled Dot-Product Attention to highlight critical sentiment tokens.

3. Feedforward Layers:

a. Include 2 dense layers, each with 128 neurons and ReLU activation.

4. Pooling Layer:

a. Perform mean pooling over token embeddings for sentence representation.

5. Output Layer:

a. Fully connected softmax layer for multi-class sentiment classification.

Step 6: Model Training

Input: Training dataset D_train, hyperparameters H

Output: Trained model M

1. Initialize training hyperparameters:

a. Batch size = 32.

b. Learning rate = 2e-5 with dynamic scheduling.

c. Dropout rate = 0.3.

d. Loss function = Cross-Entropy Loss.

e. Optimizer = Adam**.**

2. Split dataset D_train:

a. Perform k fold cross validation with k=5

b. Use randomization in the selection of the components of each and every fold

3. Train model M for 100 epochs:

a. Monitor validation loss.

b. Stop training early if validation loss does not improve for 3 consecutive epochs (early stopping).

Step 7: Model Testing and Evaluation

Input: Testing dataset D_test, trained model M

Output: Evaluation metrics

1. Evaluate M on D_test:

a. Compute APS from accuracy, Cohen's kappa and Matthews Correlation Coefficient (MCC).

b. Analyze error cases, focusing on misclassified implicit negations.

2. Document patterns in errors and potential areas for augmentation or model improvement.

Step 8: Validation

Input: Augmented dataset D_augmented, model M

Output: Validated model performance

1. Perform 5-fold cross-validation on D_augmented.

2. Analyze performance consistency across folds.

3. Use statistical tests to ensure results are significant. END.

5.5 Discussion

The development of the hybrid deep learning model leads to a successful achievement of the third objective and reveals compelling insights. It is evidenced from the theoretical analysis that majority of the existing sentiment analysis models detects a minimal number of the polarity shifters leading to inaccurate sentiment classification. It is clear from both theoretical and empirical analysis that back translation is a powerful data augmentation technique. Based on this, addressing polarity shift through advancements in translation models, context-aware approaches, quality checks, and domain adaptation strategies can further enhance back translation's effectiveness in handling polarity shift and improving sentiment analysis model's performance on diverse datasets.

The newly developed deep hybrid model has addressed the challenge of handling polarity shift due to implicit negation through enhancing back translation by preserving specific context and idiomatic expressions. For example, the Amazon reviews dataset often contain colloquial language, domain- specific terms, and nuanced sentiments that are crucial for accurate sentiment analysis and polarity shift management. It is evidenced that back translation may not always capture these aspects effectively leading to a loss of context and potentially affecting model performance. The newly developed hybrid model has improved the back translation data augmentation through implementation of perturbations by introducing controlled variations in the back-translated data while preserving important contextual and idiomatic elements. In the newly developed hybrid model, perturbations preserve specific context and idiomatic expressions during back translation, such as, the idiom retention, domain specific terms and the cultural sensitivity. Incorporating domain-specific dictionaries or lexicons to guide back translation and ensure that important terms related to mobile devices or product features are accurately translated is critical. This helps in creating a diverse training dataset for the hybrid deep learning model, enhancing its ability to handle polarity shift and nuanced sentiments. Perturbations were also combined with the back translation data augmentation strategy to enrich the training data with diverse linguistic patterns and sentiment expressions.

Therefore the newly developed hybrid model enhanced sentiment classification through contextual back translation augmented by Seq2seq perturbations. By combining Seq2seq models with perturbation techniques during back translation helped in creating a more diverse and contextually relevant augmented datasets for training the model enhancing their performance. This is in line with the observation made by Xu and Wang [177]. The proposed model performs better in terms of handling subtle shift that occurs when negation is implied when compared with other existing transformer based architectures. This is evidenced by the works of Ilmawan et al.[178]. Additionally, it is clear that the other transformer models do not generalize well to the scope of negations and context-dependent [188].

5.6 Summary

In this chapter, the systematic methodology for the development of the hybrid model which entails the model architecture design and development is discussed. The hybrid model leverages advanced data augmentation techniques, contextual embeddings, and attention mechanisms to enhance sentiment classification accuracy. The model development followed a structured approach that ensured methodological rigor and reproducibility. The approach incorporated data preprocessing, feature engineering process, training and hyperparameter tuning. The model architecture design is discussed showing various components of the hybrid model. The Seq2Seq perturbations and contextual back translation was incorporated to enrich the training data diversity while preserving semantic integrity. The embedding process incorporated FastText and BERT embeddings to capture both subword-level and deep contextual representations. The core architecture of the hybrid model is centered on a transformer-based and LSTM framework with an attention mechanism. The attention mechanism selectively focuses on sentiment-critical tokens, and ensuring that implicit negations are effectively identified and processed. Therefore this study contributes to the existing works through enhancement of the data augmentation process through seq2seq perturbations, contextual embedding and attention mechanism. This ensures that implicit negations are effectively identified and processed hence improved sentiment classification.

CHAPTER SIX

VALIDATION AND COMPARATIVE EVALUATION OF THE HYBRID MODEL

6.1 Introduction

This chapter presents a systematic approach to assess the effectiveness of the hybrid deep learning in detecting and classification of the implicit negation. Validation ensured that the developed hybrid model met its intended objectives of detecting and classification of the implicit negation of the mobile phone reviews. Comparative evaluation, on the other hand, evaluated the hybrid deep learning against three datasets with different variations. The performance of hybrid deep learning model was compared with four existing sentiment analysis models using both qualitative and quantitative metrics. Firstly, the details of the validation approach employed, including the four sentiment analysis models, three datasets and evaluation metrics are discussed. Both the qualitative and quantitative evaluation metric used to measure performance and to confirm the model's alignment with the hypothesis are discussed. Secondly, the chapter presents the comparative evaluation, where the results of the hybrid model are compared with those of existing sentiment analysis models. This comparison highlights the relative strengths, weaknesses, and contributions of the hybrid model offering deeper insight into its practical implications.

6.2 Hybrid Model Validation Approach

The validation of the model was done to check for errors and misclassifications of the implicit negations. The experimental data analysis was adopted in order to comprehensively evaluate the performance of hybrid model on mobile phone review datasets. In this research work, model validation process involved performance of the 5-fold cross-validation on the augmented dataset in order to ensure consistency in the model’s performance across three mobile phone datasets. The classification theory underpins the model validation by ensuring minimal misclassification of the implicit negations. The empirical rigor of this approach is reinforced through comprehensive performance evaluations, employing both qualitative and quantitative evaluation metrics. The use of statistical tests is also done to ensure the results from deep hybrid model have a statistical significant difference from the existing state of the art models. The hybrid model’s performance is compared with four state of the

art models against three data sets. Both the quantitative and qualitative evaluation metrics are performed in order to ensure generalizability of the novel model. The quantitative metrics includes; Kruskal Wallis H statistics, ANOVA, average performance score (APS) that the average performance of Accuracy, Cohens Kappa and Mathews correlation coefficient and Mean Squared Error(MSE). Qualitative Metrics used are attention heat maps, whisker plots and SHAP visualization**.** The three metrics were useful because they ensured detailed analysis of models behavior, especially in tasks that involved seq2seq. The performance of the hybrid model was compared with four state of the art sentiment analysis models against three datasets of different variations.

The selection of the four sentiment analysis models used for the validation was guided by their state-of-the-art performance. This is evidenced in the theoretical analysis and the results of the pilot study as well as in the works of Murithi et al [124] and Murithi et al [139]. The four sample sentiment analysis models include: SVM for model M1, BERT and NB for model M2, BERT and LSTM for model M3, and SVM and NB for model M4. The datasets for the validation were selected through a purposive sampling. These included the Amazon Mobile Phone Reviews, Google Play Store Reviews, and iPhone Apple App Store Reviews. The choice of this datasets offered a diverse user-generated content that enabled a comprehensive evaluation of sentiment analysis models. The three key stages of validation are shown in the figure 6.1.

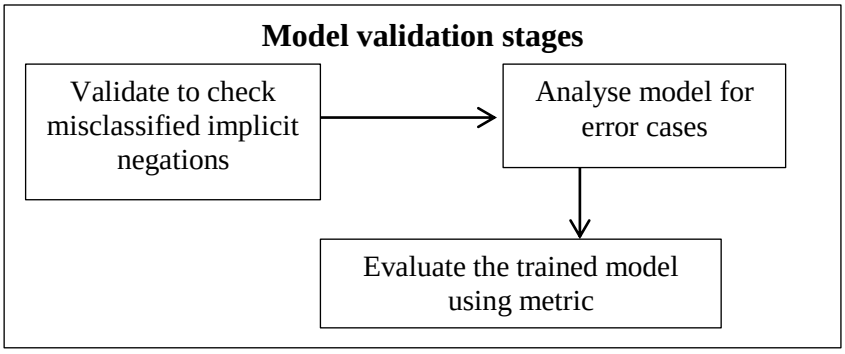


Figure 6.1 Model Validation Stages

During the validation stage, check for misclassified implicit negations was done inorder to observe any errors that stemmed from the model's inability to detect nuanced forms of negation. The transformer based component of the hybrid model as context aware capability that allowed for capturing of nuanced sentiments. However some misclassified implicit negations were detected which highlights the importance of including edge cases in training and validation datasets to improve real-world robustness. While the overall validation results demonstrated promising performance, a closer examination of the error cases was done. These error cases highlight scenarios where the hybrid model did not perform as expected. Lastly, a number of evaluation metrics was performed on the hybrid model and the validation results are discussed in the subsequent section.

6.3 Hybrid Model Validation Tests Results

This section presents the results obtained from the validation process aimed at assessing the effectiveness of the hybrid model. The evaluation was conducted based on the qualitative and quantitative metrics and criteria aligned with the research question outlined in chapter one. The validation results offer insights into how well the proposed hybrid model performs in a detecting the implicit negation. The results of the validation process are discussed in the sections 6.3.1 to 6.3.7

6.3.1 Average Performance Score of the Newly Developed Sentiment Analysis Model against the Existing Ones

In light of this, the accuracy, Cohen's kappa and Mathew's correlation coefficient metrics are performed on the newly developed hybrid model against the existing sentiment classification models. The choice of these metrics in the evaluation process is in line with the systematic review works by Murithi et al. [124].These scores are first normalized before computing their average score. The three metrics are combined in order to achieve a comprehensive evaluation of the model's performance. The accuracy metric measures how well the model predicts the sentiments into positive, negative or neutral polarities. Figure 5.4 shows the average performance score of the novel model compared to four other state of the art models. All the models were compared against three mobile phone datasets of different variability.

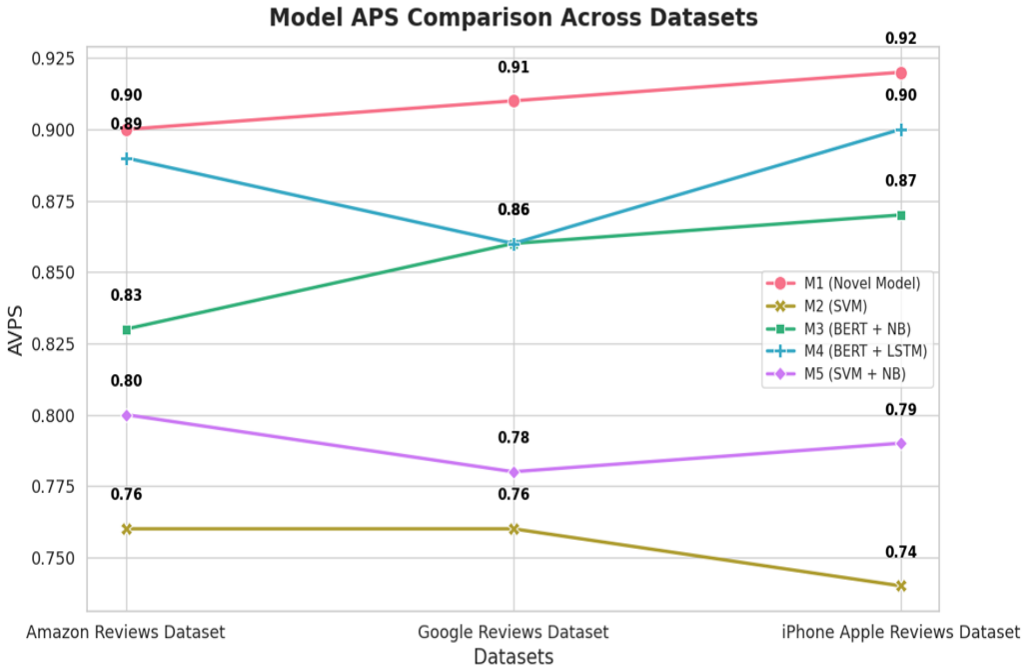


Figure 6.2: Average Performance Score of the Newly Developed Sentiment Analysis Model against the Existing Ones

M1 (Novel Model) is the most consistently high-performing model across all datasets, showing that it is the best suited for handling implicit negation. This is evident from its performance value of 0.89 on the amazon reviews dataset, 0.91 on the google reviews dataset and 0.90 on the iPhone Apple reviews dataset. The scores are notably higher than all other models in each dataset. Its primary strength lies in its ability to effectively identify and process negations within sentiment-bearing phrases, leading to more accurate sentiment predictions. M2 (SVM) consistently underperforms across all datasets, which aligns with the expectation that SVM, may struggle with more complex linguistic patterns such as implicit negation unless specifically trained on data that captures such patterns. M3 (BERT + NB) shows strong performance, but its results are slightly lower than M4, indicating that while BERT helps, Naive Bayes' simplicity might not be fully adequate for capturing the complexities of negation. M4 (BERT + LSTM) also performs very well, particularly in D1 and D3, where its ability to model both context (via BERT) and sentence structure (via LSTM) helps capture the nuances of negation. However, it slightly trails M1, suggesting that while powerful, the BERT + LSTM combination might not be as optimized for negation as the novel model. M5 (SVM + NB) performs moderately, suggesting that while this hybrid approach has its merits, it does not

fully address the challenges posed by implicit negation because it does not capture the deep contextual relationships that more advanced models like BERT.

6.3.2 The ANOVA Analysis of the Average Normalized Score of the Newly Developed

The table 6.1 shows the results of the ANOVA analysis on the average performance scores of the novel model against the existing sentiment classification models.

Table 6.1: ANOVA of the APS of the Developed Hybrid Model

ANOVA – APS

	Sum of Squares	df	Mean Square	F	p
Model	0.05084	4	0.0127	53.0	< .001
Residual	0.00240	10	2.40e-4		

The p-value of less than 0.001 exhibits statistical significant differences in the newly developed hybrid model compared with the existing sentiment analysis models. This explains that the model performs better the other four models.

Further post-hoc is done on the different state of the art models and the newly developed models and the results are shown in table 6.2.

Table 6.2: PostHoc analysis of the APS of the developed Hybrid model

Model		Model	Mean Difference	SE	df	t	p_{tukey}
M1	-	M2	0.1567	0.0126	10.0	12.39	< .001
	-	M3	0.0567	0.0126	10.0	4.48	0.008
	-	M4	0.0267	0.0126	10.0	2.11	0.288
	-	M5	0.1200	0.0126	10.0	9.49	< .001
M2	-	M3	-0.1000	0.0126	10.0	-7.91	< .001

	-	M4	-0.1300	0.0126	10.0	-10.28	< .001
	-	M5	-0.0367	0.0126	10.0	-2.90	0.092
M3	-	M4	-0.0300	0.0126	10.0	-2.37	0.200
	-	M5	0.0633	0.0126	10.0	5.01	0.004
M4	-	M5	0.0933	0.0126	10.0	7.38	< .001

Note. Comparisons are based on estimated marginal means

The post hoc analysis shows statistical significant difference between Novel model (M1) and M2, M3, and M5 with values of <0.01, 0.08 and <0.01 respectively. M2 has some statistical significance difference with M3 and M4 with values of <0.001. There are also notable statistical significance differences between M3 and M5 with values of 0.04 as well as between M4 and M5 with values of <0.001. This clearly indicated that the model performs better than the existing state of the art models**.**

6.3.3 Kruskal Wallis H Statistic of the Newly Developed Sentiment Analysis Model against the Existing Ones

Kruskal-Wallis H test is used for validating the newly developed hybrid model as it is a non-parametric method ideal for comparing median ranks across multiple groups. This test is key for this research work because it is suitable for non-homogeneity of variance which applies to all real world data. Table 6.3 shows the results of the Kruskal wallis H statistic. It is clear from the results that there is a significance difference between the newly developed novel model and the existing sentiment analysis model with a p value of 0.011.

Table 6.3 Kruskal-wallis Results of the APS of the Hybrid Model

Kruskal-Wallis (NON PARAMETRIC)

	χ^2	df	p	ε^2
APS	**13.0**	**4**	**0.011**	**0.931**

Further post Hoc analysis was carried out and W reflects the degree of difference between groups. It is clear that the models M1 (Novel Model) shows a higher ranking compared to M2, M3, M4 and M5 with values ranging from -2.40 to -2.82. The results of the groups of the various models are shown in table 5.4.

Table 6.4 Dwass -Steel- Critchlow-Fligner Pairwise Comparisons

Pairwise comparisons – APS

		W	p
M1	M2	-2.82	0.270
M1	M3	-2.78	0.284
M1	M4	-2.40	0.391
M1	M5	-2.78	0.284
M2	M3	2.82	0.270
M2	M4	2.82	0.270
M2	M5	2.82	0.270
M3	M4	1.88	0.674
M3	M5	-2.78	0.284
M4	M5	-2.78	0.284

6.3.4 Mean Squared Error of the Newly Developed Sentiment Analysis Model against the Existing Ones

Mean squared error (MSE) was used to measure the difference between predicted values (x) and true values (y). The smaller the MSE, the better the model's prediction align with the true values. M1 (Novel Model) is the most consistently high-performing model with MSE values of 0.06, 0.07 and 0.08 across the three datasets. This is evident from its performance, which is notably higher than all other models.

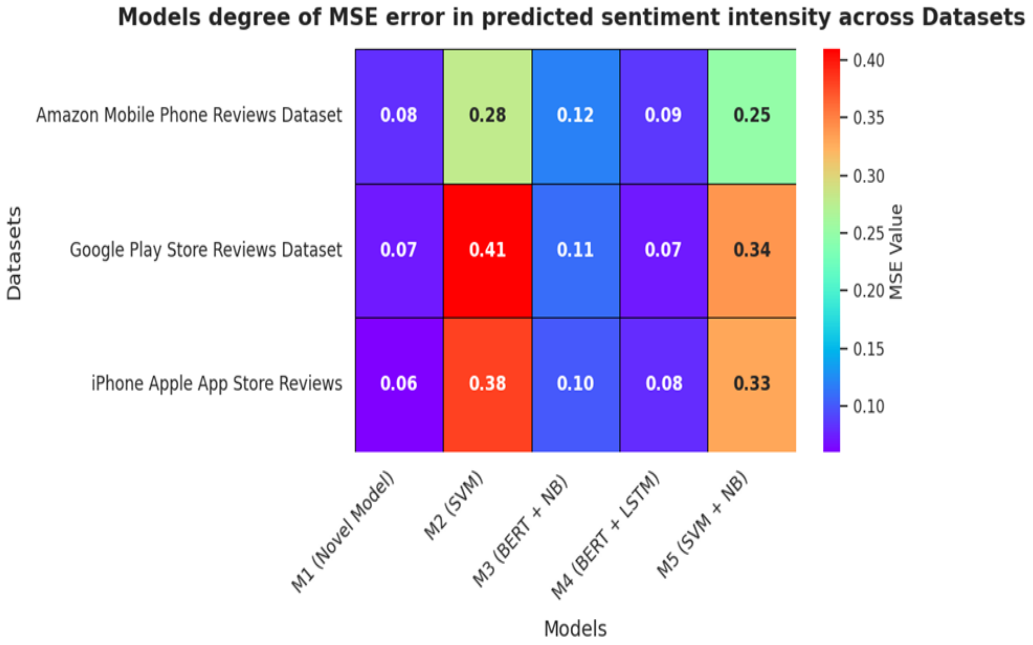


Figure 6.3: Mean Squared Error of the Newly Developed Model against the Existing Ones

The sentiment analysis models, designed to detect polarity shifts caused by implicit negation, demonstrate varying levels of effectiveness across the datasets. In the context of sentiment analysis, lower MSE (Mean Squared Error) values indicate better model performance because MSE measures the difference between the predicted sentiment values and the actual sentiment labels. A lower MSE reflects smaller errors, meaning the model's predictions are closer to the true sentiment.

The process enhances the optimization theory by adopting optimization of loss function in order to minimize errors.

In specific:

i. M1 (Novel Model) performs exceptionally well across all datasets, with MSE values between 0.06 and 0.08. These low values suggest that this model effectively captures the nuanced shifts in sentiment caused by implicit negation, making it the most accurate model in this comparison.

ii. M2 (SVM) shows much higher MSE values (ranging from 0.28 to 0.41), indicating that it struggles to correctly predict sentiment in the presence of implicit negation. The higher the MSE, the larger the gap between predicted and actual sentiment, pointing to poorer performance.

iii. M3 (BERT + NB) and M4 (BERT + LSTM) both exhibit relatively low MSE values, signifying solid performance. M4 (BERT + LSTM) slightly outperforms M3 (BERT + NB) in most cases, making it a better choice for handling implicit negation in sentiment analysis.

iv. M5 (SVM + NB) shows moderate performance with MSE values that are higher than M1, M3, and M4, but not as high as M2, indicating that while it handles polarity shifts reasonably well, it is not the most effective model across all datasets.

6.3.5 Whisker Plots of the Newly Developed Sentiment Analysis Model against the Existing Ones

The whisker plot shows relatively lower median (second quartile) values for the sentiment analysis models M2, M3, M4 and M5 as compared to the median values of the newly developed hybrid model. The whiskers for the newly developed hybrid model M1 are relatively shorter than the ones in model M2, M3, M4 and M5. The hybrid model demonstrates superior performance and reliability compared to the existing sentiment analysis models. Higher median values indicate better average performance scores, while the shorter whiskers and lack of outliers suggest more consistent and reliable implicit sentiment classification.

Figure 6.4 shows the whisker plots of the average performance score of the newly developed hybrid model and existing sentiment classification models.

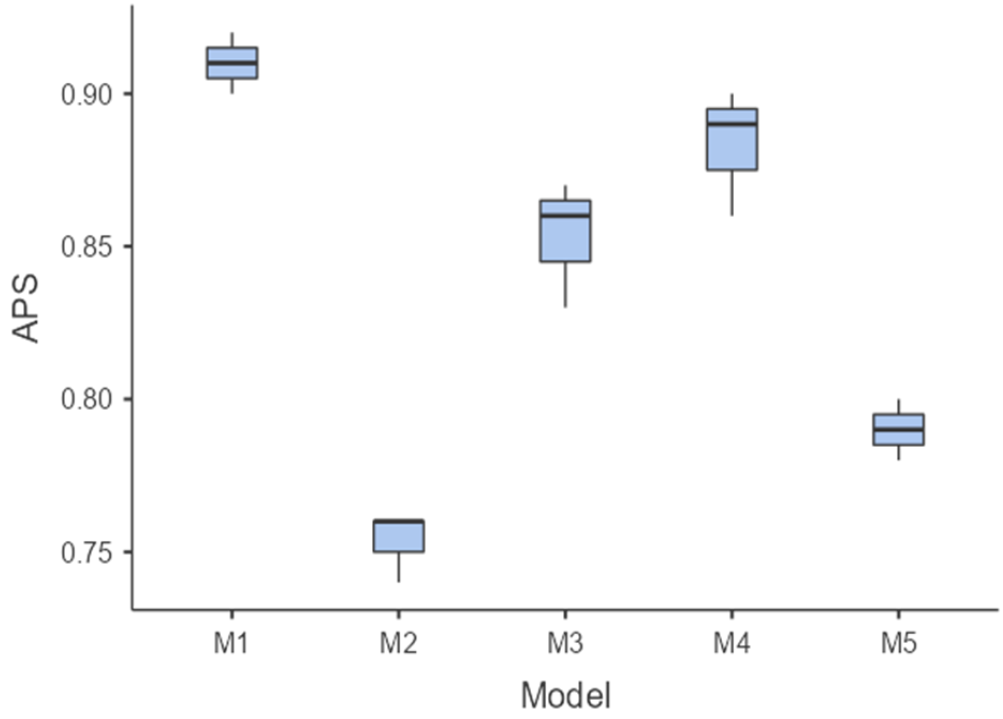


Figure 6.4: Whisker Plots of the Newly Developed Sentiment Analysis Model against the Existing Ones

6.3.6 Attention Heat Maps of the Newly Developed Sentiment Analysis Model against the Existing Ones.

The attention heatmap visualizes the performance of five sentiment analysis models (M1, M2, M3, M4, M5) across three datasets using the plasma colour palette. In this research work the heatmap, bright yellow indicates the highest average performance score, while dark purple represents the lowest. It is clear in results presented in figure 5.9 that the M1 (Novel Model) consistently performs well across all datasets, with bright yellow shades indicating high average performance score.

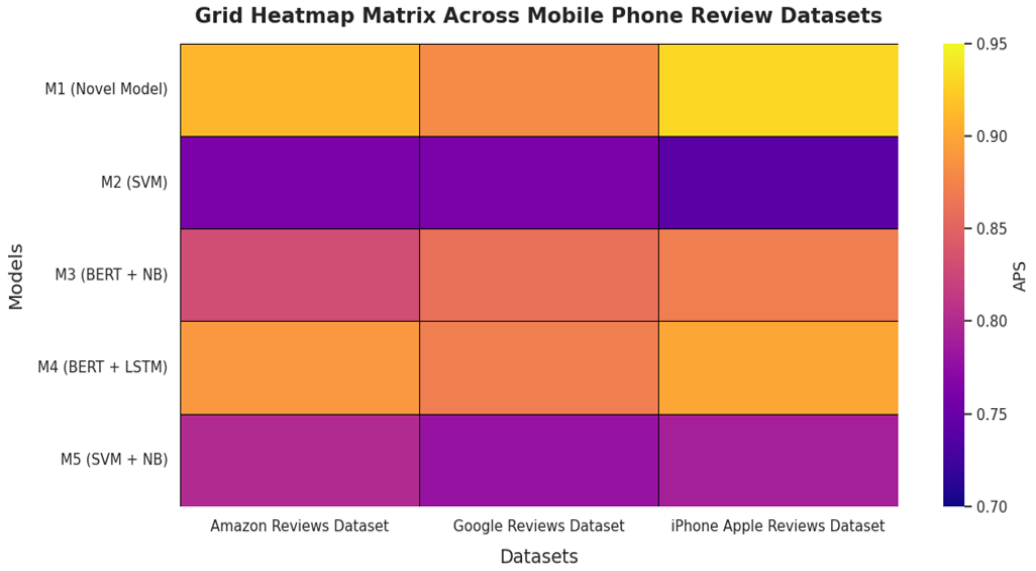


Figure 6.5: Attention Heat Maps of the Newly Developed Sentiment Analysis Model against the Existing Ones

The heatmap visualizes the performance of five sentiment analysis models (M1, M2, M3, M4, M5) across three datasets (Amazon Reviews, Google Reviews, and iPhone Apple Reviews) using the **plasma** colour palette. In this heatmap, **bright yellow** indicates the highest average performance score, while **dark purple** represents the lowest. The **M1 (Novel Model)** consistently performs well across all datasets, with bright yellow shades indicating high average performance score, making it the top-performing model. **M4 (BERT + LSTM)** also shows strong performance, with orange and yellow colours, particularly excelling in the iPhone Apple Reviews dataset (D3). **M3 (BERT + NB)** demonstrates moderate performance, marked by a mix of red and yellow shades, while **M5 (SVM + NB)** shows slightly lower average performance score, as indicated by the orange and red hues. **M2 (SVM)** consistently performs the worst, with purple/red shades across all datasets, indicating it struggles with handling polarity shifts due to implicit negation. The heatmap effectively uses the plasma colour scheme to highlight the strengths and weaknesses of each model, making it easy to identify which models are most effective at managing implicit negation across different mobile review datasets.

6.3.7 Hybrid Model Performance Visualization using SHAP

Polarity shifts, where the sentiment of a word or phrase changes due to implicit negation, pose significant challenges in the explainability of machine learning models used for sentiment classification. Incorporating mechanisms like sentiment score distributions provides a more granular view of sentiment by evaluating specific components, offering explainable insights into the model's sentiment classification accuracy[189];[190]. Adjusting for these shifts can improve transparency and classification accuracy. A machine learning approach that accounts for polarity-shifting information has been shown to enhance sentiment classification accuracy and transparency. In a study focusing on sentiment classification transparency in polarity-shifting detection, researchers proposed a machine learning approach that fully considers the polarity-shifting phenomenon[191]. They first extracted detection rules to identify polarity shifts and then incorporated this information into the sentiment classification process. This method demonstrated explainability in sentiment classification tasks. Furthermore, a systematic review of sentiment analysis models highlighted the importance of explainability in models' detection of polarity shifts [192]. The review analysed various models and their explainability in managing polarity shifts, emphasizing that incorporating mechanisms to models' explainability in the detection of these shifts can significantly enhance model transparency.

Figures 6.6 and 6.7 demonstrate the explainability of the hybrid deep learning model across one of the mobile phone reviews datasets. SHAP visualization is used in this research study because it helped visualize which features contributed most to positive or negative sentiment prediction.

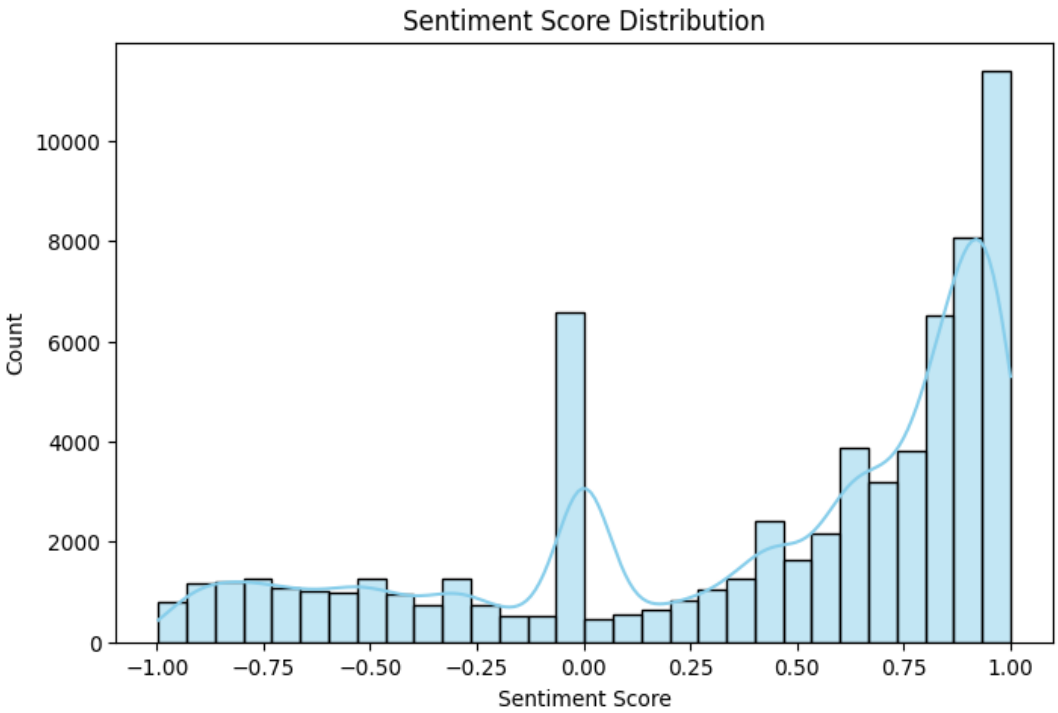


Figure 6.6: Sentiment Score Distribution of the Combined Title and Body Features in Amazon Mobile Phone Datasets

Figure 5.10 illustrates the distribution of sentiment scores in amazon mobile phones reviews datasets, highlighting how the model captures subtle polarity shifts. The distribution has combined the title and the body features which are key for sentiment analysis. The probability theory was applied in the allocation of the sentiment scores whereby -1 to -0.5 indicates a negation polarity. The score of -0.5 to 0.5 indicates a neutral polarity while 0.5 to 1 indicates a positive polarity. It is clear from the figure that most of the reviews are skewed to the positive values which indicates positive polarity. About 60,000 reviews indicate the neutral polarity. Further visualization is performed to compare the impact of the title and body of the sentiment individually in the sentiment classification. The result is shown in figure 6.7.

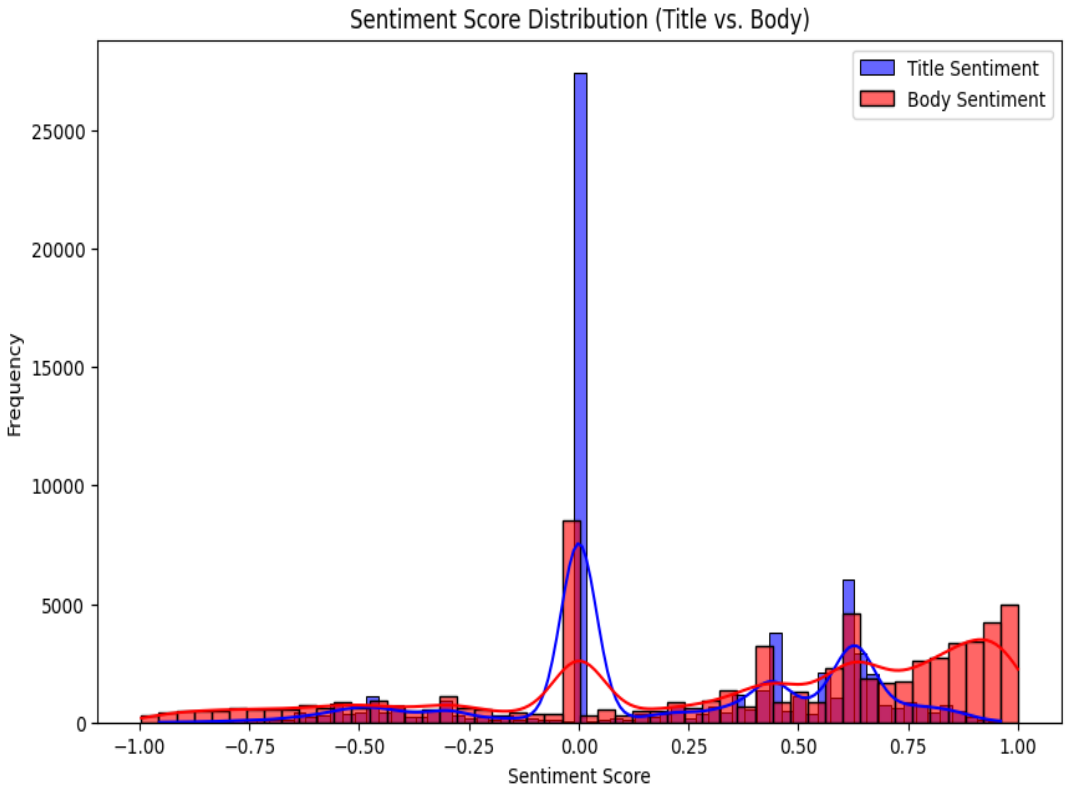


Figure 6.7: Sentiment Score Distribution of Title and Body Features in Amazon Mobile Phone Reviews Datasets

Figure 6.7 showcases the sentiment score distribution in amazon mobile phone reviews datasets emphasizing the model's ability to detect implicit negations. This figure presents most of the sentiment score distribution skewed to the positive side indicating a positive polarity when the body of the sentiment is used for the sentiment classification. It is clear that over 25000 reviews were skewed to the neutral polarity when the title of the sentiment is used which is contrary when the body of the sentiment is used. This indicates that the body of the sentiment is key in the hybrid model during the sentiment classification. These results demonstrate the model's effectiveness in handling complex polarity shifts contained in the body of the sentiments. These SHAP visualizations underscore the model's enhanced explainability and accuracy in sentiment classification tasks involving polarity shifts.

The average sentiment score was computed using the formula below:

$$\mu = \frac{\sum f x i}{\sum f} \qquad \text{Average sentiment score}$$

Where

μ = the mean or average.

$\sum$ = Summation symbol ("sum of...").

f = Frequency of each value (how many times it occurs).

x_i = the value (or midpoint, in grouped data) corresponding to that frequency.

$\sum f$ = Total number of observations (sum of all frequencies).

$\sum f \cdot x_i$ = the sum of each value multiplied by its frequency.

6.4 Model Validation Triangulation

Triangulation is a research strategy that integrates multiple methods to enhance the validity and reliability of results findings [193]. This study adopted triangulation in the validation of the hybrid model. The validation triangulation in this study was achieved through the use of multiple metrics, three datasets with different variability to ensure generazability and four existing state of the art models for comparison. Both the quantitative and qualitative evaluation metrics were adopted in order to enhance the validity and reliability of newly developed hybrid deep learning model. The quantitative metrics includes; Kruskal Wallis H statistics, ANOVA, Accuracy, Cohens Kappa, Mathews correlation coefficient and Mean Squared Error(MSE). Qualitative Metrics used are attention heat maps, whisker plots and SHAP visualization. The statistical results from the ANOVA and kruskal wallis H statistical analysis show that the newly developed model has statistical significance difference compared with the existing state of the art models. The statistical results are in agreement with the results from attention heat maps and whisper plots. The model was validated using three mobile phone datasets of different variability. The hybrid model achieved average performance score of 0.89 on the amazon reviews dataset, 0.91 on the google reviews dataset and 0.90 on the iPhone Apple reviews dataset. These scores are notably higher than the other four existing models in each dataset. The insights derived from this the model validation triangulation contribute to advancing the field of sentiment analysis, with implications extending beyond mobile phone reviews to broader domains where sentiment interpretation.

6.5 Error Analysis of the Hybrid Deep Learning Model

The newly developed hybrid deep learning model was validated using both qualitative and quantitative metrics against three datasets to ensure generalizability. The model achieved an average performance score of above 0.89 across all the three datasets used. The newly developed sentiment analysis model consistently performs high across all datasets, showing that it is the best suited for handling implicit negation. However, some failure cases from the hybrid model where implicit negation was misclassified were identified through performance of the mean squared error analysis. The results of the mean squared error shows that the hybrid model performs exceptionally well across all datasets, with MSE values between 0.06 and 0.08 compared to the state of the art models. These low values suggest that this model effectively captures the nuanced shifts in sentiment caused by implicit negation. However, the hybrid model suffered some failure cases where implicit negation was misclassified. A qualitative error analysis was performed on the hybrid model and a table that shows a sample of misclassified sentiments of amazon mobile phone reviews dataset was created. The results are shown in table 6.5.

Table 6.5: Qualitative Error Analysis of the Hybrid Model Based on a Sample of Amazon Mobile Phone Reviews

Row Number	**Title**	**Body**	**Model's sentiment score & classification**	**Ground Truth Sentiment**
67981	Updated review	Update: If you like sending and receiving picture texts regularly you might want to wait until the bugs get worked out.	0.48 (Neutral)	Positive
589	used phone still the way to go	These are pretty bullet proof phones…. Not so much the battle scars.	-0.51 (Negative)	Neutral
8092	Great Phone	I researched the crap out of this….. Much easier than evernote!	0.52 (Positive)	Neutral

19796	Excellent phone, the best Android...	Let me preface this by saying I am a device and tech haund... but I expected better build quality.	0.49 (Neutral)	Positive
1367	Samsung focus I917	Pro: amazing AMOLED screen display. Really beautiful. Cons: Slow performance, feels outdated.	-0.45 (Neutral)	Negative

Table 6.5 have a sampled the amazon mobile phone reviews that were misclassified by the hybrid model. The main causes of this misclassification was due to nuanced sentiment where the model detects positive attributes but hesitates to classify the review as fully positive, interpreting the cautious phrasing and qualifying statements as signals of partial reservation. Additionally the model struggles to balance the nuanced criticisms with mixed sentiments and contextual contradictions, leading to a misclassification. The possible caused for deviation are explained.

In row 589 the possible causes for deviation of ground truth vs. computed sentiment score from the model is due to its limited ability to interpret implicit negation and contextual sentiment. Although the reviewer acknowledges the phone's durability with phrases like "bulletproof phones" and confirms its functionality despite wear, the model likely overemphasized the implicit dissatisfaction conveyed through indirect complaints. Phrases such as "long wait," "flat battery," and "battle scars" suggest negative experiences without explicitly using harsh or emotional language. The model struggled to balance these nuanced criticisms against the positive aspects, leading to a misclassification. This reflects a common challenge in sentiment analysis accurately interpreting subtle or implied negative cues within an otherwise neutral or mixed context.

In row 67981 possible causes for deviation of ground truth vs. computed sentiment score from the model is due to its ability to detect implicit negation and mixed

sentiment. While the reviewer praises features like the 4K OLED display, Dolby Atmos sound, and game enhancer, they also express persistent frustrations with SMS glitches, fingerprint reader issues, and game exits. Phrases such as "If you like sending and receiving picture texts regularly, you might want to wait until the bugs get worked out" convey reservations despite overall enthusiasm. The model recognizes how positive descriptions are repeatedly undermined by these ongoing problems, leading to a more balanced sentiment classification.

In row 8092 possible causes for misclassification could have arisen due to the hybrid model's inability to detect implicit negation embedded within the review. Although the review mentions positive aspects like "a big improvement" and "the Verizon deal was perfect," there are subtle negative implications hidden in statements like "I did not purchase this phone through Amazon," and "this is a preliminary review." These phrases diminish the certainty and positive sentiment by implying unverified satisfaction or potential future issues. The model overlooks these indirect negations, leading to an incorrect positive classification instead of the more accurate neutral sentiment.

In row 19796 possible causes for deviation of ground truth vs. computed sentiment score from the model is due to the presence of implicit negation embedded within the positive language. While the review emphasizes the phone's exceptional build quality, high performance, and premium feel, subtle comparisons and qualifiers introduce a level of restraint that tempers the overall sentiment. For instance, the phrase "Not that the S5 was bad but..." implicitly acknowledges that the previous device was adequate, downplaying the absolute superiority of the new phone. Additionally, the reviewer mentions that the stylus functions well "even though I use an Otterbox," implying a potential limitation or concern that is mitigated but still present. This results in the classification leaning toward strong neutral towards positive rather than outright positive.

In row 1367 possible causes for hybrid model classified the text as neutral inclined towards negative, as opposed to outright negative, due to the presence of implicit negation, which dampens the positive tone. Although the reviewer praises aspects like the AMOLED display, smooth handling, and responsive speed, these positives are consistently followed by subtle limitations. For instance, the statement "works

better with AT&T" implies the phone functions inadequately on other networks. Similarly, "battery drains fast" suggests a major inconvenience for heavy users. The update mentioning that the phone "sometimes will quit starting" conveys a serious reliability issue but is presented in a neutral, factual manner rather than an emotionally negative tone. These implicit negatives, despite the positive points, lead the model to classify the sentiment as neutral leaning negative rather than outright negative.

6.6 Discussion

The newly developed hybrid deep learning model was validated with a number of quantitative metrics that include accuracy, Cohen’s kappa, Mathews correlation and mean squared error. The average performance score that comprised accuracy, Cohens kappa, Mathews correlation showed consistently high-performance across all datasets, showing that it is the best suited for handling implicit negation compared to the four other existing models. The newly developed hybrid model performs well across all datasets, with MSE values between 0.06 and 0.08. These low values suggest that this model effectively captures the nuanced shifts in sentiment caused by implicit negation, making it the most accurate model in this comparison. Statistical analyses, including the Kruskal-Wallis H statistic and subsequent ANOVA tests, underscored significant differences in performance, highlighting the superior efficacy of the new developed hybrid deep learning model across three datasets. The p-value of less than 0.001 exhibits statistical significant differences in the newly developed hybrid model compared with the existing sentiment analysis models. The whisper plots also demonstrate higher median values and the shorter whiskers and lack of outliers suggesting a more consistent and reliable implicit sentiment classification.

Lastly, it is clear from attention heat map results that the newly developed model consistently performs well across all datasets, with bright yellow shades indicating high average performance score, making it the top-performing model. Notably, leveraging on the contextual back translation augmented by Seq2seq perturbations yield improved implicit negation detection. This is also evinced by improved sentiment classification accuracy compared with four other existing models across the three datasets. The error rate due to domain specific issues can be reduced further

by improving the training process through neural network regularization, batch normalization and transfer learning. This has been proposed as recommendations for future works in the area of deep learning. Additionally, the sentiment score demonstrated by the SHAP visualization demonstrates the model's enhanced explainability and accuracy in sentiment classification.

6.7 Hybrid Model's Limitations and Future Improvements

The hybrid deep learning model struggled in its ability to manage mixed sentiments and contextual contradictions. One promising approach is implementing a dual-stream attention mechanism, which separates positive and negative sentiment cues and evaluates them independently. Additionally, incorporating contrastive learning, a technique that trains the model to distinguish between similar but opposing sentiment patterns, will improve the model's ability to recognize implicit contradictions. This is especially useful when a review contains praise alongside reservations, as the model can learn to classify both elements without allowing one to dominate the sentiment evaluation. Future research could also explore contextual sentiment disentanglement, allowing the model to treat positive and negative aspects as independent components while still accounting for how they interact to shape overall sentiment.

The hybrid model had some difficulty in detecting indirect dissatisfaction embedded within neutral or positive statements. Future work should focus on improving context-aware negation detection. This involves enhancing the model's ability to interpret subtle negations (e.g., "not so much the battle scars") that do not follow explicit negation patterns but soften or reverse the sentiment. Implementing multi-task learning where the model simultaneously predicts sentiment polarity and negation presence can significantly boost performance. This dual focus will allow the model to recognize both positive language and underlying dissatisfaction, reducing the likelihood of misclassifying mixed or ambiguous reviews. Future improvements should also consider adaptive contextual embedding, which dynamically weighs the contribution of positive and negative sentiment cues.

The hybrid model's struggle with hedging language and preliminary statements can be mitigated by incorporating causal inference models that can differentiate between speculative and definitive statements. Such models analyse the causal relationship between sentiment cues and how qualifiers (e.g., "I did not purchase through Amazon") affect sentiment certainty. By focusing on the causal flow of sentiment, the model can better interpret reviews that indicate potential future dissatisfaction without immediately classifying them as positive. Additionally, synthetic data augmentation techniques should be employed to create training samples that contain uncertain phrasing (e.g., "preliminary review") and qualified statements.

The hybrid model struggled with accurately classifying sentiment in cases with coexisting positive and negative aspects. The model was unable to disentangle different aspects of a product (e.g., build quality, performance) and assign a sentiment polarity to each. In reviews where positive statements (e.g., "excellent phone") are paired with critical observations (e.g., "I expected better build quality"), the model can capture these dual sentiments separately rather than averaging them into a neutral classification. Another enhancement involves discourse-level attention, which tracks sentiment across multiple sentences. This would prevent the model from overweighting isolated negative phrases or dismissing positive affirmations.

The hybrid model struggled to classify sentiment in reviews where major concerns are tempered by positive descriptions. Future research should focus on hierarchical sentiment modelling. This method processes sentiment at multiple levels by first identifying primary and secondary sentiment themes and then assigning a weight based on their relative importance. Future work could also explore saliency-based sentiment extraction, where the model learns to identify and prioritize emotionally charged phrases over neutral statements. This improvement would help the model avoid misclassification when subtle negative cues (e.g., "feels outdated") play a more critical role than surface-level praise.

6.8 Summary

The validation process incorporated both qualitative and quantitative metrics. To validate accuracy, Cohen kappa, Mathews correlation and mean squared error metrics were used across three datasets compared with four state of the art models. The model was validated across three datasets. The results of the hybrid model demonstrate increased sentiment classification accuracy in its ability to handle implicit negation. Further ANOVA and kruskal wallis H statistics tests are used to test for the statistical significance of the hybrid models against the existing models. This test aimed at testing for the hypothesis stated in chapter one. The ANOVA and kruskal wallis H statistics results of $p<0.01$ showed that there is statistical significance difference between the deep hybrid model and the existing models for handling sentiment analysis. There the Null Hypothesis stated in chapter was rejected. Lastly, the visualization of hybrid model results using the attention heat maps and whisker plots showed its efficacy in handling implicit negations with improved accuracy. The SHAP visualization improved the hybrid model interpretability by evaluating title and body of the sentiment as the components offering insights into the model's sentiment classification accuracy. This validation process ensures that the hybrid model effectively detects implicit negations and hence improved sentiment classification.

CHAPTER SEVEN

SUMMARY, CONCLUSION AND RECOMMENDATIONS

7.1 Summary

Sentiment analysis main aim is to analyse the users' sentiments and classify them as positive, neutral or negative. Thus, the classification theory of machine learning serves as the foundation for sentiment analysis. To ensure accurate sentiment classification the model should be able to detect and handle the polarity shifters present in the sentiments. This study has addressed the polarity shift detection in sentiment analysis therefore improving the sentiment classification accuracy. The study developed a hybrid deep learning model that detects the implicit negation polarity shifters. Since the implicit negations are not explicitly expressed, the existing sentiment analysis model tends to ignore them hence misclassifying the sentiments. It is notable that even with the recent advancement of deep learning and the use of pre-trained embeddings and attention mechanism in sentiment analysis, the challenge of detecting implicit negation remains an open area of research.

Firstly, in this study, the review of the literature in sentiment analysis models used in handling polarity shift is carried out in order to identify the key polarity shifters in written text. The review encompasses a critical evaluation and summary of the existing sentiment analysis models used in handling polarity shift as well as feature selection and extraction techniques used with these models. The datasets used and metrics for evaluating the performance of these models is also discussed. The literature review results establish that the main challenge facing sentiment analysis is polarity shift. It is clear from the review that negation is the most significant polarity shifter in sentiment analysis. Even with the advancement of deep learning implicit negation, domain-specific linguistic nuances, and dynamic shifts in sentiment expressions are identified as the challenges. The conceptual framework guides the empirical research work of this study.

The second objective aimed to conduct an empirical research in order to investigate the performance of the various machine learning models used in handling the polarity shift due to implicit negation in a variety of feature selection, feature extraction and data augmentation techniques. The empirical analysis process provides invaluable insights into both technical and theoretical research

contributions of this thesis. The theoretical implications of our findings extend beyond model development, offering a deeper understanding of the interplay between different techniques in the context of sentiment analysis at the face of implicit negation. Experiment on the state of the art models selected during a pilot study is conducted. Three variables including feature selection, feature extraction and data augmentation techniques are used with the models and their interaction compared using amazon mobile phone dataset. The average performance score that comprises of accuracy, Cohens kappa and Mathew's correlation are the metrics used for the comparison of the models' performance. The models' training process was optimized through back translation data augmentation and hyper parameter tuning and the various performance metrics compared. ANOVA analysis is done to test for the statistical significance difference of the different models and the interaction of the various variables used in the experiments. The results show that there is an interaction between the deep learning models and the data augmentation technique but no interactions with feature selection techniques as well as feature extraction techniques. This is because feature selection and extraction happen in the deep network of the deep learning models.

The third objective aimed at the developing a hybrid deep learning model for handling polarity shift problem due to implicit negation in sentiment analysis of the mobile phone brands' reviews. The model aimed to improve sentiment classification through detection of the implicit negation present in the sentiments. This is achieved by incorporating FastText and BERT embedding's ability to detect nuanced polarity shift by capturing both subword and deep contextual representation. A structured approach for experiments set up was adopted that incorporates data preprocessing, feature engineering, model training and hyperparameter training. Data preprocessing has been achieved through a number of steps. BERT's word tokenizer and standard tokenization handles any variations. Text normalization for all the text is done. Lastly data augmentation process leverages on the contextual back translation augmented by Seq2seq perturbations inorder to enrich the training the process. Feature extraction and selection are the two feature engineering processes performed. A transformer based BERT model with attention mechanism is incorporated in the hybrid model inorder to selectively focus on the sentiment crucial tokens. This ensures implicit negations are effectively selected. The training process

leverages on an adaptive learning rate and scheduling and drop out regularization. The two strategies optimize the convergence and generalization performance. The hybrid deep learning model addresses the identified challenges by integrating advanced techniques to improve sentiment analysis accuracy, robustness, and contextual understanding in mobile phones reviews.

The fourth objective sought to analyse the model's performance against the state of the art sentiment analysis models. Both qualitative and quantitative metrics are performed on the hybrid model to ensure generalizability. Kruskal Wallis H statistics, ANOVA, average performance score (APS) and Mean Squared Error are the quantitative metrics used. Attention Heat Maps, whisker plots and SHAP visualization are the three metrics used for qualitative evaluation. The hybrid model performed better against the other state of the art models against three data sets. The ANOVA and kruskal wallis H statistics demonstrated that the newly developed hybrid model has a statistical significance difference compared with the existing models. The SHAP visualization demonstrates the model's interpretability and it was clear from the results that the title and the body features are two aspect of the dataset that had more impact on the model's performance. The performance metrics demonstrates that the hybrid model improves the sentiment classification accuracy.

7.2 Conclusion

Polarity shift is one of the main challenges in sentiment analysis that leads to misclassifying of the sentiment and hence wrong decision making by the companies. This research study therefore addresses the implicit negation polarity shifter by developing a hybrid deep learning model. The results of this research can effectively be used to improve the brand reputation through improvement of the features and customer service delivery through accurate classification of customers' sentiments. Recent advancements in NLP have introduced models incorporating deep learning, pre-trained embeddings, and attention mechanisms. This study therefore leverages on LSTM deep learning model and transformer based BERT model to detect the implicit negation in mobile phone sentiment analysis. The sentiment analysis model leverages on the contextual back translation augmented by Seq2seq perturbations inorder to enrich the training process. The model incorporates FastText and BERT embedding inorder to detect nuanced polarity shift by capturing both subword and

deep contextual representation. An attention mechanism was adopted inorder to selectively focus on sentiment-critical tokens, and ensures that implicit negations are identified and processed. The training process ensures effective differentiation between sentiment classes through adoption of Cross-Entropy Loss, which is ideal for multi-class classification tasks. The training process optimizes the loss function hence minimizing the errors during the sentiment classification.

The study identifies negation as one of the main polarity shifters in sentiment analysis. It clear that addressing the complexities of negation in sentiment analysis models is crucial for enhancing their overall dynamism and effectiveness in managing polarity shifts. The existing sentiment analysis models struggle to address the negation polarity shifter because they primarily rely on explicit negation markers (e.g., 'not,' 'never'), hence fail to recognize context-dependent sentiment reversals. This therefore affects the classification accuracy, leading to misinterpretations in consumer feedback, brand reputation analysis, and automated decision-making. This research study therefore has addressed the implicit negation in mobile phone review sentiments. The newly developed hybrid deep model has enhanced the back translation by preserving specific context and idiomatic expressions. This is evidenced from the Amazon mobile phone reviews dataset used that often contain colloquial language, domain- specific terms, and nuanced sentiments.

Experiments on the state of the art sentiment analysis models selected during the pilot study shows that there are interactions among the traditional machine learning models and feature selection and feature extraction but none with data augmentation. Therefore it is very key to utilize a feature selection and extraction techniques that optimizes the performance of the traditional machine learning model in sentiment analysis. Deep learning interacts effectively with the data augmentation techniques, therefore incorporating one during sentiment analysis, improves the sentiment classification. Back translation data augmentation performs well in sentiment analysis but encounters the challenge context retention. It is evidenced that back translation faces challenge of not always been able to preserve the specific context, or idiomatic expressions. This study has adopted context-aware back translation approaches that consider the context of the original text and try to preserve idiomatic expressions or domain-specific language. Poor translations introduce noise or

incorrect sentiment cues into the augmented data, affecting model performance. This study has incorporated a contextual back translation augmented by Seq2seq perturbations.

The development of the hybrid deep learning model integrates advanced data augmentation, feature handling, and architectural strategies. This aims to address the challenge of implicit negation which can shift sentiment polarity in subtle but significant ways. The model training is based on Cross-Entropy Loss and hyperparameter on Adam optimizer so as to ensure robustness, adaptability, and high performance. The data preprocessing implements a data augmentation strategy that combines contextual back translation and Seq2Seq perturbations. The seq2seq perturbations create controlled syntactic variations, ensuring semantic integrity while enriching the training data with contextually relevant and linguistically diverse examples. FastText embeddings and BERT embeddings are incorporated in the feature engineering process. FastText captures semantic relationships at the subword level. BERT embeddings provide deep contextual representations, enabling the model to understand complex linguistic structures and nuances in negation. The hybrid deep learning model demonstrates improved performance compared to the existing state of the art sentiment classification models. The average performance score metrics indicates that the newly developed model outperforms the existing models with 0.97 accuracy. The newly developed hybrid model shows a mean squared error (MSE) of 0.08 with the amazon mobile reviews datasets, 0.07 with google play store data sets and 0.06 with Iphone app reviews datasets compared to the existing models which has as high as 0.41. This is a clear indication that the low mean square error indicates better model performance. The result of the attention heat maps indicates that the newly developed model consistently performs well across all datasets. The shorter whiskers and lack of outliers suggest more consistent and reliable implicit sentiment classification of the newly developed model. The results of ANOVA and kruskal wallis H statistics indicates that there is a significance difference between the newly developed hybrid deep learning model and the existing sentiment analysis model. The researcher therefore rejects the alternate hypothesis and accepts the null hypothesis. The evaluation of the newly developed model against three datasets ensures detailed analysis of models behaviour hence improvement of the model's generalizability. This research work

has minimized misclassification of the sentiments hence improving the sentiment classification accuracy.

7.3 Recommendation on Policy and Practice

This research study proposes the following recommendations on policy and practice.

7.3.1 Integration with Business and Decision Systems

The hybrid deep learning model demonstrates its ability to detect implicit negation and hence improve sentiment classification accuracy. This result contributes in improved decisions making process with direct applications in customer feedback analysis across sectors such as finance, hospitality, and healthcare. The hybrid deep learning model can be deployed in the business and decision systems to enhance the decision making process. In these domains informed decisions based on customer opinions, reviews, and interactions are made. Furthermore, this research holds implications for brand monitoring, social media sentiment analysis, and market intelligence, where companies rely on sentiment data to shape strategies, understand consumer preferences, and improve services. The enhanced model's ability to handle implicit negation ensures more accurate sentiment classification, leading to more reliable insights in these critical areas.

7.3.2 Testing with Contextual Embedding for Implicit Negation

The study incorporates FastText and BERT embedding inorder to detect nuanced polarity shift by capturing both subword and deep contextual representation. This model’s architecture can implement attention-based deep learning models that capture subtle negation. Additionally, models that use domain-specific fine-tuning can be implemented using this architecture inorder to enhance model sensitivity to negation in industry-specific language.

7.3.3 Testing with Hybrid Model Design

The hybrid deep model leverages advanced data augmentation techniques, contextual embeddings, and attention mechanisms to enhance sentiment classification accuracy. Contextual back translation augmented by Seq2seq perturbations enhances the training process. The BERT embeddings provide deep contextual representations, enabling the model to understand complex linguistic structures and nuances in negation. The hybrid model design can be used inorder to effectively detect polarity

shifters, dependency parsing, and contextual embeddings negation in complex sentences.

7.3.4 Adoption of Comprehensive Evaluation Metrics

The hybrid deep learning model was evaluated with both qualitative and quantitative performance metrics to ensure detailed analysis of models behaviour. The model was evaluated across three dataset in order to ensure consistency in the model's performance. The use of statistical tests i.e. ANOVA and Kruskal wallis H statistics in this model ensures the results have a statistical significant difference from the existing state of the art models. The novel model's performance is compared with four state of the art models against three data sets. The comprehensive evaluation metrics and use of statistical significance test can be adopted to evaluate sentiment analysis models. SHAP visualization can be adopted as well inorder to improve sentiment models' interpretability.

7.4 Recommendations for Future Research

This study has incorporated a contextual back translation augmented by Seq2seq perturbations and achieved improved sentiment classification accuracy. Implicit negation detection has been achieved through the adoption of FastText embeddings and BERT embeddings in the feature engineering process. This captures semantic relationships at the subword level, making it effective for handling spelling variations and domain-specific terms often found in reviews. However, there are issues of model architecture, handling of multimodal sentiments and optimization of training process that this study recommends for future works.

7.4.1 Experiments with other Training Optimization Approaches

This study recommends that in future a model training process that aim at exploring the other approaches that can handle data scarcity and imbalances in real world other than data augmentation techniques can be implemented. In sentiment analysis approached like neural network regularization, batch normalization and transfer learning are some of the approaches that can be explored. The improved training process can also create training samples that contain uncertain phrasing therefore handling the challenge of hedge language and preliminary statements during sentiment classification.

7.4.2 Experiments with Computation Resource Efficiency Environment

This study recommends future research directions of optimizing computational efficiency through knowledge distillation, pruning, and quantization. These avenues can enhance the model's performance while reducing computational resource requirements. This will also enable real-time sentiment analysis in resource-constrained environments.

7.4.3 Experiment with Multimodal Sentiment Analysis

Further studies can focus on the combination of other datasets like audio and video analysis in order to understand polarity shifts that include sarcasm and tone changes in speech. Additionally, facial expressions and voice modulations can be explored to detect hidden sentiment inversions in social media sentiments and customer feedback. This will ensure human-level sentiment understanding is achieved.

7.4.4 Experiment with Mixed Sentiments and Contextual Contradictions

This study recommends use of contrastive learning to distinguish between similar but opposing sentiments. Contextual sentiment disentanglement is a challenge whereby the positive sentiment and negative sentiments are not treated individually hence leading to contradictions. Additionally, to handle this challenge, the study recommends use of dual-stream attention mechanism can be implemented to separate positive and negative cues and evaluate them individually.

REFERENCES

[1] A. Kafi, M. S. Ashikul Alam, S. Bin Hossain, S. B. Awal, and H. Arif, "Feature-Based Mobile Phone Rating Using Sentiment Analysis and Machine Learning Approaches," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh: IEEE, May 2019, pp. 1–6. doi: 10.1109/ICASERT.2019.8934555.

[2] A. M. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, Chennai, India: IEEE, Jan. 2017, pp. 72–76. doi: 10.1109/ICoAC.2017.7951748.

[3] K. Khan, B. Baharudin, A. Khan, and A. Ullah, "Mining opinion components from unstructured reviews: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 26, no. 3, pp. 258–275, Sep. 2014, doi: 10.1016/j.jksuci.2014.03.009.

[4] M. Tubishat, N. Idris, and M. A. M. Abushariah, "Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges," *Inf. Process. Manag.*, vol. 54, no. 4, pp. 545–563, Jul. 2018, doi: 10.1016/j.ipm.2018.03.008.

[5] M. Bordoloi and S. K. Biswas, "Sentiment analysis: A survey on design framework, applications and future scopes," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 12505–12560, Nov. 2023, doi: 10.1007/s10462-023-10442-2.

[6] Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, "Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training," Nov. 03, 2021, *arXiv*: arXiv:2111.02194. Accessed: Sep. 26, 2024. [Online]. Available: http://arxiv.org/abs/2111.02194

[7] J. Miao and L. Niu, "A Survey on Feature Selection," *Procedia Comput. Sci.*, vol. 91, pp. 919–926, 2016, doi: 10.1016/j.procs.2016.07.111.

[8] M. Tubishat, N. Idris, L. Shuib, M. A. M. Abushariah, and S. Mirjalili, "Improved Salp Swarm Algorithm based on opposition based learning and novel local search algorithm for feature selection," *Expert Syst. Appl.*, vol. 145, p. 113122, May 2020, doi: 10.1016/j.eswa.2019.113122.

[9] M. Wankhade, A. C. S. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.

[10] S. A. Parvin, M. Sumathi, and C. Mohan, “Challenges of Sentiment Analysis - A Survey,” in *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India: IEEE, Jun. 2021, pp. 781–786. doi: 10.1109/ICOEI51242.2021.9453026.

[11] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, “Natural Language Processing Advancements By Deep Learning: A Survey,” 2020, *arXiv*. doi: 10.48550/ARXIV.2003.01200.

[12] J. Zhou, J. Tian, R. Wang, Y. Wu, W. Xiao, and L. He, “SentiX: A Sentiment-Aware Pre-Trained Model for Cross-Domain Sentiment Analysis,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 568–579. doi: 10.18653/v1/2020.coling-main.49.

[13] M. Van De Kauter, D. Breesch, and V. Hoste, “Fine-grained analysis of explicit and implicit sentiment in financial news articles,” *Expert Syst. Appl.*, vol. 42, no. 11, pp. 4999–5010, Jul. 2015, doi: 10.1016/j.eswa.2015.02.007.

[14] A. S. Abuhammad and M. A. Ahmed, “Negation Detection Techniques in Sentiment Analysis: A Survey,” *Iraqi J. Sci.*, pp. 1060–1069, Feb. 2024, doi: 10.24996/ijs.2024.65.2.37.

[15] A. Kumar and G. Garg, “Empirical study of shallow and deep learning models for sarcasm detection using context in benchmark datasets,” *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 5, pp. 5327–5342, May 2023, doi: 10.1007/s12652-019-01419-7.

[16] W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[17] Y. Mao, Q. Liu, and Y. Zhang, “Sentiment analysis methods, applications, and challenges: A systematic literature review,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.

[18] K. Pipalia, R. Bhadja, and M. Shukla, “Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis,” in *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*, Moradabad, India: IEEE, Dec. 2020, pp. 411–415. doi: 10.1109/SMART50582.2020.9337081.

[19] P. Sudheer, J. Manoranjini, N. Suman, E. M. Reddy, and P. Sridevi, “A Comprehensive Exploration of Complex Emotions and Their Simplification for Enhancing Sentiment Analysis Through Explainable AI,” in *Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 1*, vol. 1273, A. Kumar, V. K. Gunjan, S. Senatore, and Y.-C. Hu, Eds., in Lecture Notes in Electrical Engineering, vol. 1273. , Singapore: Springer Nature Singapore, 2025, pp. 975–985. doi: 10.1007/978-981-97-8031-0_103.

[20] R. Satapathy, Y. Li, S. Cavallari, and E. Cambria, “Seq2Seq Deep Learning Models for Microtext Normalization,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest, Hungary: IEEE, Jul. 2019, pp. 1–8. doi: 10.1109/IJCNN.2019.8851895.

[21] M. D. Devika, C. Sunitha, and A. Ganesh, “Sentiment Analysis: A Comparative Study on Different Approaches,” *Procedia Comput. Sci.*, vol. 87, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[22] M. U. Salur and I. Aydin, “A Novel Hybrid Deep Learning Model for Sentiment Classification,” *IEEE Access*, vol. 8, pp. 58080–58093, 2020, doi: 10.1109/ACCESS.2020.2982538.

[23] R. Xia, F. Xu, J. Yu, Y. Qi, and E. Cambria, “Polarity shift detection, elimination and ensemble: A three-stage model for document-level sentiment analysis,” *Inf. Process. Manag.*, vol. 52, no. 1, pp. 36–45, Jan. 2016, doi: 10.1016/j.ipm.2015.04.003.

[24] N. S. Mohd Nafis and S. Awang, “An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification,” *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[25] N. K. Singh, D. S. Tomar, and A. K. Sangaiah, “Sentiment analysis: a review and comparative analysis over social media,” *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 1, pp. 97–117, Jan. 2020, doi: 10.1007/s12652-018-0862-8.

[26] C. I. Eke, A. A. Norman, and L. Shuib, “Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model,” *IEEE Access*, vol. 9, pp. 48501–48518, 2021, doi: 10.1109/ACCESS.2021.3068323.

[27] M. Yadav and R. Katarya, “Polarity Shift Detection and Handling in Polarity Classification: A Survey,” *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3562967.

[28] R. Kaur and S. Kautish, “Multimodal Sentiment Analysis: A Survey and Comparison,” in *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, I. R. Management Association, Ed., IGI Global, 2022, pp. 1846–1870. doi: 10.4018/978-1-6684-6303-1.ch098.

[29] P. Korenek and M. Šimko, “Sentiment analysis on microblog utilizing appraisal theory,” *World Wide Web*, vol. 17, no. 4, pp. 847–867, Jul. 2014, doi: 10.1007/s11280-013-0247-z.

[30] H. Mohanty, “NLP for Sentiment Computation,” in *Trends of Data Science and Applications*, vol. 954, S. S. Rautaray, P. Pemmaraju, and H. Mohanty, Eds., in Studies in Computational Intelligence, vol. 954. , Singapore: Springer Singapore, 2021, pp. 1–21. doi: 10.1007/978-981-33-6815-6_1.

[31] M. Taboada, “Sentiment Analysis: An Overview from Linguistics,” *Annu. Rev. Linguist.*, vol. 2, no. 1, pp. 325–347, Jan. 2016, doi: 10.1146/annurev-linguistics-011415-040518.

[32] S. Ghosh, A. Ekbal, and P. Bhattacharyya, “Natural language processing and sentiment analysis: perspectives from computational intelligence,” in *Computational Intelligence Applications for Text and Sentiment Data Analysis*, Elsevier, 2023, pp. 17–47. doi: 10.1016/B978-0-32-390535-0.00007-0.

[33] P. Etingof *et al.*, “Introduction to representation theory,” Feb. 01, 2011, *arXiv*: arXiv:0901.0827. doi: 10.48550/arXiv.0901.0827.

[34] D. Tang, “Sentiment-Specific Representation Learning for Document-Level Sentiment Analysis,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai China: ACM, Feb. 2015, pp. 447–452. doi: 10.1145/2684822.2697035.

[35] K. P. Gunasekaran, “Exploring Sentiment Analysis Techniques in Natural Language Processing: A Comprehensive Review,” 2023, doi: 10.48550/ARXIV.2305.14842.

[36] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, “Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review,” *Nat. Lang. Process. J.*, vol. 6, p. 100059, Mar. 2024, doi: 10.1016/j.nlp.2024.100059.

[37] A. Tiwari, “Supervised learning: From theory to applications,” in *Artificial Intelligence and Machine Learning for EDGE Computing*, Elsevier, 2022, pp. 23–32. doi: 10.1016/B978-0-12-824054-0.00026-5.

[38] Z. Liu, Y. Lin, and M. Sun, Eds., *Representation Learning for Natural Language Processing*. Singapore: Springer Nature Singapore, 2023. doi: 10.1007/978-981-99-1600-9.

[39] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep Learning Based Feature Selection for Remote Sensing Scene Classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015, doi: 10.1109/LGRS.2015.2475299.

[40] I. Salehin and D.-K. Kang, “A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain,”

Electronics, vol. 12, no. 14, p. 3106, Jul. 2023, doi: 10.3390/electronics12143106.

[41] K. Vayadande, H. Baru, A. Kashid, A. Kulkarni, P. Londhe, and A. Vanjari, "Text Analysis for Information Retrieval Using NLP," in *Innovations in VLSI, Signal Processing and Computational Technologies*, vol. 1095, G. Mehta, N. Wickramasinghe, and D. Kakkar, Eds., in Lecture Notes in Electrical Engineering, vol. 1095. , Singapore: Springer Nature Singapore, 2024, pp. 585–599. doi: 10.1007/978-981-99-7077-3_56.

[42] N. C. Dang, M. N. Moreno-García, and F. De La Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020, doi: 10.3390/electronics9030483.

[43] Tata Consultancy Services, Senior Solution Architect, USA and G. K. Patra, "A Sentiment Analysis of Customer Product Review Based on Machine Learning Techniques in E-Commerce," *J. Artif. Intell. Cloud Comput.*, vol. 2, no. 4, pp. 1–4, Dec. 2023, doi: 10.47363/JAICC/2023(2)389.

[44] M. Sivakumar and S. R. Uyyala, "Aspect-based sentiment analysis of mobile phone reviews using LSTM and fuzzy logic," *Int. J. Data Sci. Anal.*, vol. 12, no. 4, pp. 355–367, Oct. 2021, doi: 10.1007/s41060-021-00277-x.

[45] N. M. Alharbi, N. S. Alghamdi, E. H. Alkhammash, and J. F. Al Amri, "Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews," *Math. Probl. Eng.*, vol. 2021, pp. 1–10, May 2021, doi: 10.1155/2021/5536560.

[46] Md. N. Imtiaz and Md. K. Ben Islam, "Identifying Significance of Product Features on Customer Satisfaction Recognizing Public Sentiment Polarity: Analysis of Smart Phone Industry Using Machine-Learning Approaches," *Appl. Artif. Intell.*, vol. 34, no. 11, pp. 832–848, Sep. 2020, doi: 10.1080/08839514.2020.1787676.

[47] H.-T. Duong and T.-A. Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis," *Comput. Soc. Netw.*, vol. 8, no. 1, p. 1, Dec. 2021, doi: 10.1186/s40649-020-00080-x.

[48] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Syst. Appl.*, vol. 223, p. 119862, Aug. 2023, doi: 10.1016/j.eswa.2023.119862.

[49] N. Garg and K. Sharma, "Text pre-processing of multilingual for sentiment analysis based on social network data," *Int. J. Electr. Comput. Eng. IJECE*, vol. 12, no. 1, p. 776, Feb. 2022, doi: 10.11591/ijece.v12i1.pp776-784.

[50] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, and E. A. Retta, "Improving Arabic Sentiment Analysis Using CNN-Based Architectures and Text Preprocessing," *Comput. Intell. Neurosci.*, vol. 2021, no. 1, p. 5538791, Jan. 2021, doi: 10.1155/2021/5538791.

[51] M. Arief and M. B. M. Deris, "Text Preprocessing Impact for Sentiment Classification in Product Review," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, Jakarta, Indonesia: IEEE, Nov. 2021, pp. 1–7. doi: 10.1109/ICIC54025.2021.9632884.

[52] S. S. Kumar and A. Rajini, "Extensive Survey on Feature Extraction and Feature Selection Techniques for Sentiment Classification in Social Media," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India: IEEE, Jul. 2019, pp. 1–6. doi: 10.1109/ICCCNT45670.2019.8944391.

[53] R. Hegde and Seema S., "Aspect based feature extraction and sentiment classification of review data sets using Incremental machine learning algorithm," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Chennai, India: IEEE, Feb. 2017, pp. 122–125. doi: 10.1109/AEEICB.2017.7972395.

[54] N. S. Mohd Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," *IEEE Access*, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/ACCESS.2021.3069001.

[55] S. Alam and N. Yao, “The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis,” *Comput. Math. Organ. Theory*, vol. 25, no. 3, pp. 319–335, Sep. 2019, doi: 10.1007/s10588-018-9266-8.

[56] E. Haddi, X. Liu, and Y. Shi, “The Role of Text Pre-processing in Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.

[57] H.-T. Duong and T.-A. Nguyen-Thi, “A review: preprocessing techniques and data augmentation for sentiment analysis,” *Comput. Soc. Netw.*, vol. 8, no. 1, p. 1, Dec. 2021, doi: 10.1186/s40649-020-00080-x.

[58] D. Kurniawan, M. Yasir, and F. C. Venna, “Optimization of Sentiment Analysis using Naive Bayes with Features Selection Chi-Square and Information Gain for Accuracy Improvement,” in *2022 9th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Jakarta, Indonesia: IEEE, Oct. 2022, pp. 153–160. doi: 10.23919/EECSI56542.2022.9946510.

[59] D. Goularas and S. Kamis, “Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data,” in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*, Istanbul, Turkey: IEEE, Aug. 2019, pp. 12–17. doi: 10.1109/Deep-ML.2019.00011.

[60] P. Pongthanoo and W. Songpan, “Feature Selection and Reduction Based on SMOTE and Information Gain for Sentiment Mining,” in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, Shanghai, China: IEEE, May 2020, pp. 109–114. doi: 10.1109/ICCCS49078.2020.9118467.

[61] P. H. Prastyo, I. Ardiyanto, and R. Hidayat, “A Review of Feature Selection Techniques in Sentiment Analysis Using Filter, Wrapper, or Hybrid Methods,” in *2020 6th International Conference on Science and Technology (ICST)*, Yogyakarta, Indonesia: IEEE, Sep. 2020, pp. 1–6. doi: 10.1109/ICST50505.2020.9732885.

[62] F. Akbarian and F. Z. Boroujeni, “An Improved Feature Selection Method for Sentiments Analysis in Social Networks,” in *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, Mashhad, Iran: IEEE, Oct. 2020, pp. 181–186. doi: 10.1109/ICCKE50421.2020.9303710.

[63] M. Tubishat, M. A. M. Abushariah, N. Idris, and I. Aljarah, “Improved whale optimization algorithm for feature selection in Arabic sentiment analysis,” *Appl. Intell.*, vol. 49, no. 5, pp. 1688–1707, May 2019, doi: 10.1007/s10489-018-1334-8.

[64] S. Paudel, P. W. C. Prasad, A. Alsadoon, Md. R. Islam, and A. Elchouemi, “Feature Selection Approach for Twitter Sentiment Analysis and Text Classification Based on Chi-Square and Naïve Bayes,” in *International Conference on Applications and Techniques in Cyber Security and Intelligence ATCI 2018*, vol. 842, J. Abawajy, K.-K. R. Choo, R. Islam, Z. Xu, and M. Atiquzzaman, Eds., in Advances in Intelligent Systems and Computing, vol. 842. , Cham: Springer International Publishing, 2019, pp. 281–298. doi: 10.1007/978-3-319-98776-7_30.

[65] A. Falasari and M. A. Muslim, “Optimize Naïve Bayes Classifier Using Chi Square and Term Frequency Inverse Document Frequency For Amazon Review Sentiment Analysis,” *J. Soft Comput. Explor.*, vol. 3, no. 1, pp. 31–36, Mar. 2022, doi: 10.52465/joscex.v3i1.68.

[66] A. S. Manek, P. D. Shenoy, M. C. Mohan, and V. K. R, “Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier,” *World Wide Web*, vol. 20, no. 2, pp. 135–154, Mar. 2017, doi: 10.1007/s11280-015-0381-x.

[67] S. Kumar and S. K. Pathak, “A Novel Approach for Sentiment Analysis and Optimization Using PSO,” in *2021 International Conference on Advances in Technology, Management & Education (ICATME)*, Bhopal, India: IEEE, Jan. 2021, pp. 43–47. doi: 10.1109/ICATME50232.2021.9732752.

[68] D. Elangovan and V. Subedha, “An Effective Feature Selection Based Classification model using Firefly with Levy and Multilayer Perceptron

based Sentiment Analysis," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India: IEEE, Feb. 2020, pp. 376–380. doi: 10.1109/ICICT48043.2020.9112425.

[69] D. A. Putri, D. A. Kristiyanti, E. Indrayuni, A. Nurhadi, and D. R. Hadinata, "Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review," *J. Phys. Conf. Ser.*, vol. 1641, no. 1, p. 012085, Nov. 2020, doi: 10.1088/1742-6596/1641/1/012085.

[70] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual keyword extraction and summarization: State-of-the-art," *Inf. Process. Manag.*, vol. 56, no. 6, p. 102088, Nov. 2019, doi: 10.1016/j.ipm.2019.102088.

[71] A. Mee, E. Homapour, F. Chiclana, and O. Engel, "Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit," *Knowl.-Based Syst.*, vol. 228, p. 107238, Sep. 2021, doi: 10.1016/j.knosys.2021.107238.

[72] A. Japhne and R. Murugeswari, "Opinion Mining based complex polarity shift pattern handling for improved sentiment classification," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India: IEEE, Feb. 2020, pp. 323–329. doi: 10.1109/ICICT48043.2020.9112565.

[73] M. Ahmad and S. Aftab, "Analyzing the Performance of SVM for Polarity Detection with Different Datasets," *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 10, pp. 29–36, Oct. 2017, doi: 10.5815/ijmecs.2017.10.04.

[74] F. Alzami, E. D. Udayanti, D. P. Prabowo, and R. A. Megantara, "Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis," *Kinet. Game Technol. Inf. Syst. Comput. Netw. Comput. Electron. Control*, pp. 235–242, Aug. 2020, doi: 10.22219/kinetik.v5i3.1066.

[75] R. Hegde and Seema S., "Aspect based feature extraction and sentiment classification of review data sets using Incremental machine learning algorithm," in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics*

(AEEICB), Chennai, India: IEEE, Feb. 2017, pp. 122–125. doi: 10.1109/AEEICB.2017.7972395.

[76] S. Zirpe and B. Joglekar, "Polarity shift detection approaches in sentiment analysis: A survey," in *2017 International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, India: IEEE, Jan. 2017, pp. 1–5. doi: 10.1109/ICISC.2017.8068737.

[77] K. Thevar, V. Thakur, T. Shukla, D. Vishwakarma, and K. Ghag, "Handling polarity shift for opinion mining," in *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, Erode: IEEE, Jul. 2017, pp. 1006–1011. doi: 10.1109/ICCMC.2017.8282621.

[78] S. Kumar and S. K. Pathak, "A Novel Approach for Sentiment Analysis and Optimization Using PSO," in *2021 International Conference on Advances in Technology, Management & Education (ICATME)*, Bhopal, India: IEEE, Jan. 2021, pp. 43–47. doi: 10.1109/ICATME50232.2021.9732752.

[79] S. S. Kumar and A. Rajini, "Extensive Survey on Feature Extraction and Feature Selection Techniques for Sentiment Classification in Social Media," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kanpur, India: IEEE, Jul. 2019, pp. 1–6. doi: 10.1109/ICCCNT45670.2019.8944391.

[80] M. Avinash and E. Sivasankar, "A Study of Feature Extraction Techniques for Sentiment Analysis," in *Emerging Technologies in Data Mining and Information Security*, vol. 814, A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, and S. Dutta, Eds., in Advances in Intelligent Systems and Computing, vol. 814. , Singapore: Springer Singapore, 2019, pp. 475–486. doi: 10.1007/978-981-13-1501-5_41.

[81] P. K. Singh and S. Paul, "Deep Learning Approach for Negation Handling in Sentiment Analysis," *IEEE Access*, vol. 9, pp. 102579–102592, 2021, doi: 10.1109/ACCESS.2021.3095412.

[82] A. Kafi, M. S. Ashikul Alam, S. Bin Hossain, S. B. Awal, and H. Arif, "Feature-Based Mobile Phone Rating Using Sentiment Analysis and Machine Learning Approaches," in *2019 1st International Conference on Advances in*

Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh: IEEE, May 2019, pp. 1–6. doi: 10.1109/ICASERT.2019.8934555.

[83] B. Oscar Deho, A. William Agangiba, L. Felix Aryeh, and A. Jeffery Ansah, "Sentiment Analysis with Word Embedding," in *2018 IEEE 7th International Conference on Adaptive Science & Technology (ICAST)*, Accra: IEEE, Aug. 2018, pp. 1–4. doi: 10.1109/ICASTECH.2018.8506717.

[84] Z. Rajabi, M. R. Valavi, and M. Hourali, "A Context-Based Disambiguation Model for Sentiment Concepts Using a Bag-of-Concepts Approach," *Cogn. Comput.*, vol. 12, no. 6, pp. 1299–1312, Nov. 2020, doi: 10.1007/s12559-020-09729-1.

[85] M. Arora and V. Kansal, "Character level embedding with deep convolutional neural network for text normalization of unstructured data for Twitter sentiment analysis," *Soc. Netw. Anal. Min.*, vol. 9, no. 1, p. 12, Dec. 2019, doi: 10.1007/s13278-019-0557-y.

[86] H. Xia, C. Ding, and Y. Liu, "Sentiment Analysis Model Based on Self-Attention and Character-Level Embedding," *IEEE Access*, vol. 8, pp. 184614–184620, 2020, doi: 10.1109/ACCESS.2020.3029694.

[87] S. M. Rezaeinia, R. Rahmani, A. Ghodsi, and H. Veisi, "Sentiment analysis based on improved pre-trained word embeddings," *Expert Syst. Appl.*, vol. 117, pp. 139–147, Mar. 2019, doi: 10.1016/j.eswa.2018.08.044.

[88] R. Rajat, P. Jaroli, N. Kumar, and R. K. Kaushal, "A Sentiment Analysis of Amazon Review Data Using Machine Learning Model," in *2021 6th International Conference on Innovative Technology in Intelligent System and Industrial Applications (CITISIA)*, Sydney, Australia: IEEE, Nov. 2021, pp. 1–6. doi: 10.1109/CITISIA53721.2021.9719909.

[89] I. Irawaty, R. Andreswari, and D. Pramesti, "Vectorizer Comparison for Sentiment Analysis on Social Media Youtube: A Case Study," in *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*, Yogyakarta, Indonesia: IEEE, Sep. 2020, pp. 69–74. doi: 10.1109/IC2IE50715.2020.9274650.

[90] K. S. Kalaivani, S. Uma, and C. S. Kanimozhiselvi, “A Review on Feature Extraction Techniques for Sentiment Classification,” in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India: IEEE, Mar. 2020, pp. 679–683. doi: 10.1109/ICCMC48092.2020.ICCMC-000126.

[91] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The Impact of Features Extraction on the Sentiment Analysis,” *Procedia Comput. Sci.*, vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[92] S. Kaur, G. Sikka, and L. K. Awasthi, “Sentiment Analysis Approach Based on N-gram and KNN Classifier,” in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, Jalandhar, India: IEEE, Dec. 2018, pp. 1–4. doi: 10.1109/ICSCCC.2018.8703350.

[93] J. Mutinda, W. Mwangi, and G. Okeyo, “Lexicon-pointed hybrid N-gram Features Extraction Model (LENFEM) for sentence level sentiment analysis,” *Eng. Rep.*, vol. 3, no. 8, p. e12374, Aug. 2021, doi: 10.1002/eng2.12374.

[94] M. U. Salur and I. Aydin, “A Novel Hybrid Deep Learning Model for Sentiment Classification,” *IEEE Access*, vol. 8, pp. 58080–58093, 2020, doi: 10.1109/ACCESS.2020.2982538.

[95] E. Cambria and A. Hussain, “SenticNet,” in *Sentic Computing*, Cham: Springer International Publishing, 2015, pp. 23–71. doi: 10.1007/978-3-319-23654-4_2.

[96] A. Yadav and D. K. Vishwakarma, “Sentiment analysis using deep learning architectures: a review,” *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, Aug. 2020, doi: 10.1007/s10462-019-09794-5.

[97] S. C. Y. Chan, I. Dasgupta, J. Kim, D. Kumaran, A. K. Lampinen, and F. Hill, “Transformers generalize differently from information stored in context vs in weights,” 2022, *arXiv*. doi: 10.48550/ARXIV.2210.05675.

[98] X. Li, C. Zou, P. Wu, and Q. Li, “TSSRD: A Topic Sentiment Summarization Framework Based on Reaching Definition,” *IEEE Trans.*

Affect. Comput., vol. 14, no. 3, pp. 1716–1730, Jul. 2023, doi: 10.1109/TAFFC.2022.3186015.

[99] N. Komwad, P. Tiwari, B. Praveen, and C. R. Chowdary, "A survey on review summarization and sentiment classification," *Knowl. Inf. Syst.*, vol. 64, no. 9, pp. 2289–2327, Sep. 2022, doi: 10.1007/s10115-022-01728-y.

[100] M. Yadav and R. Katarya, "Polarity Shift Detection and Handling in Polarity Classification: A Survey," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3562967.

[101] A. Abdi, S. M. Shamsuddin, S. Hasan, and J. Piran, "Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion," *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1245–1259, Jul. 2019, doi: 10.1016/j.ipm.2019.02.018.

[102] A. Japhne and R. Murugeswari, "Opinion Mining based complex polarity shift pattern handling for improved sentiment classification," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India: IEEE, Feb. 2020, pp. 323–329. doi: 10.1109/ICICT48043.2020.9112565.

[103] N. K. Suchetha, A. Nikhil, and P. Hrudya, "Comparing the Wrapper Feature Selection Evaluators on Twitter Sentiment Classification," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, Chennai, India: IEEE, Feb. 2019, pp. 1–6. doi: 10.1109/ICCIDS.2019.8862033.

[104] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A Deeper Look into Sarcastic Tweets Using Deep Convolutional Neural Networks," 2016, doi: 10.48550/ARXIV.1610.08815.

[105] H. M. K. Kumar and B. S. Harish, "Sarcasm classification: A novel approach by using Content Based Feature Selection Method," *Procedia Comput. Sci.*, vol. 143, pp. 378–386, 2018, doi: 10.1016/j.procs.2018.10.409.

[106] P. Koncar, A. Fuchs, E. Hobisch, B. C. Geiger, M. Scholger, and D. Helic, "Text sentiment in the Age of Enlightenment: an analysis of spectator

periodicals," *Appl. Netw. Sci.*, vol. 5, no. 1, p. 33, Dec. 2020, doi: 10.1007/s41109-020-00269-z.

[107] E. Cunha, G. Magno, J. Caetano, D. Teixeira, and V. Almeida, "Fake News as We Feel It: Perception and Conceptualization of the Term 'Fake News' in the Media," in *Social Informatics*, vol. 11185, S. Staab, O. Koltsova, and D. I. Ignatov, Eds., in Lecture Notes in Computer Science, vol. 11185. , Cham: Springer International Publishing, 2018, pp. 151–166. doi: 10.1007/978-3-030-01129-1_10.

[108] W. A. Davis, *Irregular Negatives, Implicatures, and Idioms*, vol. 6. in Perspectives in Pragmatics, Philosophy & Psychology, vol. 6. Dordrecht: Springer Netherlands, 2016. doi: 10.1007/978-94-017-7546-5.

[109] S. Wang, G. Lv, S. Mazumder, and B. Liu, "Detecting Domain Polarity-Changes of Words in a Sentiment Lexicon," 2020, doi: 10.48550/ARXIV.2004.14357.

[110] Shahnawaz and P. Astya, "Sentiment analysis: Approaches and open issues," in *2017 International Conference on Computing, Communication and Automation (ICCCA)*, Greater Noida: IEEE, May 2017, pp. 154–158. doi: 10.1109/CCAA.2017.8229791.

[111] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Int. J. Surg.*, vol. 8, no. 5, pp. 336–341, 2010, doi: 10.1016/j.ijsu.2010.02.007.

[112] A. M. Abirami and V. Gayathri, "A survey on sentiment analysis methods and approach," in *2016 Eighth International Conference on Advanced Computing (ICoAC)*, Chennai, India: IEEE, Jan. 2017, pp. 72–76. doi: 10.1109/ICoAC.2017.7951748.

[113] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, and U. R. Acharya, "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis," *Future Gener. Comput. Syst.*, vol. 115, pp. 279–294, Feb. 2021, doi: 10.1016/j.future.2020.08.005.

[114] Z. Li, Y. Zou, C. Zhang, Q. Zhang, and Z. Wei, “Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training,” Nov. 03, 2021, *arXiv*: arXiv:2111.02194. doi: 10.48550/arXiv.2111.02194.

[115] J. Barnes, E. Velldal, and L. Øvrelid, “Improving sentiment analysis with multi-task learning of negation,” *Nat. Lang. Eng.*, vol. 27, no. 2, pp. 249–269, Mar. 2021, doi: 10.1017/S1351324920000510.

[116] H. Fei, Y. Ren, and D. Ji, “Negation and speculation scope detection using recursive neural conditional random fields,” *Neurocomputing*, vol. 374, pp. 22–29, Jan. 2020, doi: 10.1016/j.neucom.2019.09.058.

[117] N. Pröllochs, S. Feuerriegel, B. Lutz, and D. Neumann, “Negation scope detection for sentiment analysis: A reinforcement learning framework for replicating human interpretations,” *Inf. Sci.*, vol. 536, pp. 205–221, Oct. 2020, doi: 10.1016/j.ins.2020.05.022.

[118] S. Zirpe and B. Joglekar, “Negation Handling using Stacking Ensemble Method,” in *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, PUNE, India: IEEE, Aug. 2017, pp. 1–5. doi: 10.1109/ICCUBEA.2017.8463946.

[119] S. Mathapati, A. Nafeesa, R. Tanuja, S. H. Manjula, and K. R. Venugopal, “Semi-supervised domain adaptation and collaborative deep learning for dual sentiment analysis,” *SN Appl. Sci.*, vol. 1, no. 8, p. 907, Aug. 2019, doi: 10.1007/s42452-019-0943-0.

[120] O. Montenegro, O. S. Pabon, and R. E. G. De Pinerez R., “A Deep Learning Approach for Negation Detection from Product Reviews written in Spanish,” in *2021 XLVII Latin American Computing Conference (CLEI)*, Cartago, Costa Rica: IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/CLEI53233.2021.9640190.

[121] A. Khandelwal and S. Sawant, “NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution,” 2019, doi: 10.48550/ARXIV.1911.04211.

[122] L. Lazib, Y. Zhao, B. Qin, and T. Liu, “Negation scope detection with recurrent neural networks models in review texts,” *Int. J. High Perform. Comput. Netw.*, vol. 13, no. 2, p. 211, 2019, doi: 10.1504/IJHPCN.2019.097501.

[123] L. Chen, “Attention-Based Deep Learning System for Negation and Assertion Detection in Clinical Notes,” *Int. J. Artif. Intell. Appl.*, vol. 10, no. 01, pp. 1–9, Jan. 2019, doi: 10.5121/ijaia.2019.10101.

[124] Murithi, M.K, Oirere, A.M, and Ndung’u,R.N, “A Systematic Review of the Sentiment Analysis Models Used in Handling Polarity Shift,” *Int. J. Comput. Sci. Res.*, vol. 8, pp. 2635–2676, 2024.

[125] D. Elangovan and V. Subedha, “An Effective Feature Selection Based Classification model using Firefly with Levy and Multilayer Perceptron based Sentiment Analysis,” in *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India: IEEE, Feb. 2020, pp. 376–380. doi: 10.1109/ICICT48043.2020.9112425.

[126] A. Deniz, M. Angin, and P. Angin, “Evolutionary Multiobjective Feature Selection for Sentiment Analysis,” *IEEE Access*, vol. 9, pp. 142982–142996, 2021, doi: 10.1109/ACCESS.2021.3118961.

[127] P. Pongthanoo and W. Songpan, “Feature Selection and Reduction Based on SMOTE and Information Gain for Sentiment Mining,” in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, Shanghai, China: IEEE, May 2020, pp. 109–114. doi: 10.1109/ICCCS49078.2020.9118467.

[128] A. Rasool, R. Tao, M. Kamyab, and S. Hayat, “GAWA–A Feature Selection Method for Hybrid Sentiment Classification,” *IEEE Access*, vol. 8, pp. 191850–191861, 2020, doi: 10.1109/ACCESS.2020.3030642.

[129] J. Gorai and D. K. Shaw, “Two-phase feature extraction with chi square-based feature selection approach for the classification of online product reviews,” in *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India: IEEE, Jun. 2022, pp. 1433–1438. doi: 10.1109/ICCES54183.2022.9835934.

[130] N. P. Cruz, M. Taboada, and R. Mitkov, “A machine-learning approach to negation and speculation detection for sentiment analysis,” *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 9, pp. 2118–2136, Sep. 2016, doi: 10.1002/asi.23533.

[131] T. A. A. Abdullah, M. S. M. Zahid, W. Ali, and S. U. Hassan, “B-LIME: An Improvement of LIME for Interpretable Deep Learning Classification of Cardiac Arrhythmia from ECG Signals,” *Processes*, vol. 11, no. 2, p. 595, Feb. 2023, doi: 10.3390/pr11020595.

[132] M. L. Baptista, K. Goebel, and E. M. P. Henriques, “Relation between prognostics predictor evaluation metrics and local interpretability SHAP values,” *Artif. Intell.*, vol. 306, p. 103667, May 2022, doi: 10.1016/j.artint.2022.103667.

[133] H. J. Kwon, H. I. Koo, and N. I. Cho, “Improving Explainability of Integrated Gradients with Guided Non-Linearity,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy: IEEE, Jan. 2021, pp. 385–391. doi: 10.1109/ICPR48806.2021.9412936.

[134] I. Benbasat and R. W. Zmud, “Empirical Research in Information Systems: The Practice of Relevance,” *MIS Q.*, vol. 23, no. 1, p. 3, Mar. 1999, doi: 10.2307/249403.

[135] J. W. Creswell, *Qualitative inquiry & research design: choosing among five approaches*, 2nd ed. Thousand Oaks (Calif.): Sage publ, 2007.

[136] M. J. Page *et al.*, “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[137] J. Pallant, *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*, 7th ed. Routledge, 2020. doi: 10.4324/9781003117452.

[138] Pertti Järvinen, “various characteristics of action research,” Aug. 2009.

[139] Murithi, M.K, Oirere, A.M, and Ndung’u,R.N, “Empirical Analysis of the State-of-the-Art Models for Handling Polarity Shifts Due to Implicit Negation in Mobile Phone Reviews,” *Int. J. Comput. Sci. Res.*, vol. 1, 2025.

[140] I. Etikan, “Comparison of Convenience Sampling and Purposive Sampling,” *Am. J. Theor. Appl. Stat.*, vol. 5, no. 1, p. 1, 2016, doi: 10.11648/j.ajtas.20160501.11.

[141] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[142] G. Kaur and A. Sharma, “A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis,” *J. Big Data*, vol. 10, no. 1, p. 5, Jan. 2023, doi: 10.1186/s40537-022-00680-6.

[143] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018, *arXiv*. doi: 10.48550/ARXIV.1810.04805.

[144] C. Sun, L. Huang, and X. Qiu, “Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence,” 2019, *arXiv*. doi: 10.48550/ARXIV.1903.09588.

[145] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[146] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, “Pre-trained models for natural language processing: A survey,” *Sci. China Technol. Sci.*, vol. 63, no. 10, pp. 1872–1897, Oct. 2020, doi: 10.1007/s11431-020-1647-3.

[147] Sida Wang and Christopher D. Manning, “Baselines and Bigrams: Simple, Good Sentiment and Topic Classification,” *Proc. 50th Annu. Meet. Assoc. Comput. Linguist.*, pp. 89–93, 2012.

[148] J. McAuley and J. Leskovec, “Hidden factors and hidden topics: understanding rating dimensions with review text,” in *Proceedings of the 7th ACM conference on Recommender systems*, Hong Kong China: ACM, Oct. 2013, pp. 165–172. doi: 10.1145/2507157.2507163.

[149] B. Liu, *Sentiment Analysis and Opinion Mining*, 1st ed. 2012. in Synthesis Lectures on Human Language Technologies. Cham: Springer International Publishing, 2012. doi: 10.1007/978-3-031-02145-9.

[150] H. Xie, W. Lin, S. Lin, J. Wang, and L.-C. Yu, “A multi-dimensional relation model for dimensional sentiment analysis,” *Inf. Sci.*, vol. 579, pp. 832–844, Nov. 2021, doi: 10.1016/j.ins.2021.08.052.

[151] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene Selection for Cancer Classification using Support Vector Machines,” *Mach. Learn.*, vol. 46, no. 1/3, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

[152] S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, “Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data,” in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, Da Nang, Vietnam: IEEE, Oct. 2019, pp. 1–8. doi: 10.1109/KSE.2019.8919368.

[153] Juan Ramos, “Using TF-IDF to Determine Word Relevance in Document Queries,” *Proc. First Instr. Conf. Mach. Learn.*, vol. 242, no. 1, pp. 29–48, Dec. 2003.

[154] K. Spärck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *J. Doc.*, vol. 60, no. 5, pp. 493–502, Oct. 2004, doi: 10.1108/00220410410560573.

[155] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching Word Vectors with Subword Information,” *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.

[156] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of Tricks for Efficient Text Classification,” 2016, *arXiv*. doi: 10.48550/ARXIV.1607.01759.

[157] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Sep. 07, 2013, *arXiv*: arXiv:1301.3781. doi: 10.48550/arXiv.1301.3781.

[158] R. Sennrich, B. Haddow, and A. Birch, “Neural Machine Translation of Rare Words with Subword Units,” Jun. 10, 2016, *arXiv*: arXiv:1508.07909. doi: 10.48550/arXiv.1508.07909.

[159] J. Wei and K. Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” 2019, *arXiv*. doi: 10.48550/ARXIV.1901.11196.

[160] F. Yasser, S. Hatem, L. Ayman, L. Mohamed, and A. Mohammed, “Data Augmentation for Sentiment Analysis Enhancement: A Comparative Study,” in *2024 6th International Conference on Computing and Informatics (ICCI)*, New Cairo - Cairo, Egypt: IEEE, Mar. 2024, pp. 53–58. doi: 10.1109/ICCI61671.2024.10485144.

[161] Perry, M, Fawcett, R, and Johnson, P., “Experimental design and checklists in scientific research: A systematic approach,” *Sci. Methods Rev.*, vol. 11, no. 4, pp. 37–47, 2019.

[162] Hernandez, J and Rios, M, “A review of checklists and their applications in research.,” *J. Exp. Res.*, vol. 15, no. 2, pp. 89–101.

[163] Liu, J, Wang, T., and Zhang, Y, “A framework for effective data preprocessing in machine learning,” *Int. J. Data Sci. Mach. Learn.*, vol. 5, no. 3, pp. 22–35, 2017.

[164] Pedregosa, F *et al.*, “Scikit-learn: Machine learning in Python. the Journal of machine Learning research,” no. 12, pp. 2825–2830, 2011.

[165] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[166] Manning C.D, *Manning, C. D. (2009). An introduction to information retrieval.* Cambridge University Press, 2009.

[167] Pennington, J., Socher, R., and Manning, C. D., “Glove: Global vectors for word representation.,” *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. EMNLP*, pp. 1532–1543, Oct. 2014.

[168] Kohavi, R, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” *Ijcai*, vol. 14, no. 2, pp. 1137–1145, Aug. 1995.

[169] L. J. Cronbach, “Coefficient Alpha and the Internal Structure of Tests,” *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951, doi: 10.1007/BF02310555.

[170] J. Huang, Y. Xue, X. Hu, H. Jin, X. Lu, and Z. Liu, “Sentiment analysis of Chinese online reviews using ensemble learning framework,” *Clust. Comput.*, vol. 22, no. S2, pp. 3043–3058, Mar. 2019, doi: 10.1007/s10586-018-1858-z.

[171] H. Singhvi, “Socio-Economic Impacts of Predictive Policing on Minority Communities and Potential Solutions,” Nov. 27, 2024. doi: 10.58445/rars.2024.

[172] Sarah Shelley, “Understanding the Ethics of Data Collection and Responsible Data Usage.”

[173] M. H. I. Ahmad Hapez, N. L. Adam, and Z. Ibrahim, “Performance Analysis of Machine Learning Techniques for Sentiment Analysis,” in *Advances in Visual Informatics*, vol. 13051, H. Badioze Zaman, A. F. Smeaton, T. K. Shih, S. Velastin, T. Terutoshi, B. N. Jørgensen, H. Aris, and N. Ibrahim, Eds., in Lecture Notes in Computer Science, vol. 13051. , Cham: Springer International Publishing, 2021, pp. 205–213. doi: 10.1007/978-3-030-90235-3_18.

[174] R. Gikera, E. Maina, S. M. Mambo, and J. Mwaura, “K-hyperparameter tuning in high-dimensional genomics using joint optimization of deep differential evolutionary algorithm and unsupervised transfer learning from intelligent GenoUMAP embeddings,” *Int. J. Inf. Technol.*, Dec. 2024, doi: 10.1007/s41870-024-02279-x.

[175] F. Prefect and P. Prosser, “Empirical Algorithmics: draw your own conclusions,” Dec. 10, 2014, *arXiv*: arXiv:1412.3333. doi: 10.48550/arXiv.1412.3333.

[176] C. C. McGeoch, *A Guide to Experimental Algorithmics*, 1st ed. Cambridge University Press, 2012. doi: 10.1017/CBO9780511843747.

[177] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “mixup: Beyond Empirical Risk Minimization,” 2017, *arXiv*. doi: 10.48550/ARXIV.1710.09412.

[178] A. R. Pathak, B. Agarwal, M. Pandey, and S. Rautaray, “Application of Deep Learning Approaches for Sentiment Analysis,” in *Deep Learning-Based Approaches for Sentiment Analysis*, B. Agarwal, R. Nayak, N. Mittal, and S. Patnaik, Eds., in Algorithms for Intelligent Systems. , Singapore: Springer Singapore, 2020, pp. 1–31. doi: 10.1007/978-981-15-1216-2_1.

[179] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018, *arXiv*. doi: 10.48550/ARXIV.1810.04805.

[180] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, “A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis,” *Expert Syst. Appl.*, vol. 110, pp. 298–310, Nov. 2018, doi: 10.1016/j.eswa.2018.06.022.

[181] J. Leng and P. Jiang, “A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm,” *Knowl.-Based Syst.*, vol. 100, pp. 188–199, May 2016, doi: 10.1016/j.knosys.2016.03.008.

[182] H. Hasan, ES Qorina, K Hulliyah, A Zamhari, and D Saepudin, “Comparative Analysis of the Performance of the Fasttext and Word2vec Methods on the Semantic Similarity Query of Sirah Nabawiyah Information Retrieval System: A systematic literature review,” *8th Int. Conf. Cyber IT Serv. Manag. CITSM 2020 Virtual Oct. 23-24 2020*, Dec. 2020.

[183] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018, doi: 10.1016/j.neucom.2017.11.077.

[184] A. Onan, “Classifier and feature set ensembles for web page classification,” *J. Inf. Sci.*, vol. 42, no. 2, pp. 150–165, Apr. 2016, doi: 10.1177/0165551515591724.

[185] A. Sakai, Y. Minoda, and K. Morikawa, “Data augmentation methods for machine-learning-based classification of bio-signals,” in *2017 10th Biomedical Engineering International Conference (BMEiCON)*, Hokkaido: IEEE, Aug. 2017, pp. 1–4. doi: 10.1109/BMEiCON.2017.8229109.

[186] Lee, Michelle A., “Examining Machine Translation Systems and Translation Quality using the BackTranslation Method,” University of Wisconsin-Stout, 2021. Accessed: Jan. 13, 2025. [Online]. Available: http://digital.library.wisc.edu/1793/83072

[187] D. R. Beddiar, M. S. Jahan, and M. Oussalah, “Data expansion using back translation and paraphrasing for hate speech detection,” *Online Soc. Netw. Media*, vol. 24, p. 100153, Jul. 2021, doi: 10.1016/j.osnem.2021.100153.

[188] L. B. Ilmawan, M. Muladi, and D. D. Prasetya, “Negation handling for sentiment analysis task: approaches and performance analysis,” *Int. J. Electr. Comput. Eng. IJECE*, vol. 14, no. 3, p. 3382, Jun. 2024, doi: 10.11591/ijece.v14i3.pp3382-3393.

[189] E. Hikmawati and N. Alamsyah, “Supervised Learning for Emotional Prediction and Feature Importance Analysis Using SHAP on Social Media User Data,” *Ingénierie Systèmes Inf.*, vol. 29, no. 6, pp. 2345–2356, Dec. 2024, doi: 10.18280/isi.290622.

[190] K. Z. Yeung, F. Frasincar, and F. Van Der Knaap, “Explaining a Deep Learning Model for Aspect-Based Sentiment Analysis Using SHAP,” in *Web Information Systems Engineering – WISE 2024*, vol. 15436, M. Barhamgi, H. Wang, and X. Wang, Eds., in Lecture Notes in Computer Science, vol. 15436. , Singapore: Springer Nature Singapore, 2025, pp. 59–73. doi: 10.1007/978-981-96-0579-8_5.

[191] Y. Blázquez-López, “A Knowledge-Based Model for Polarity Shifters,” *J. Comput.-Assist. Linguist. Res.*, vol. 6, pp. 87–107, Nov. 2022, doi: 10.4995/jclr.2022.18807.

[192] T. Chaisen, P. Charoenkwan, C. G. Kim, and P. Thiengburanathum, “A Zero-Shot Interpretable Framework for Sentiment Polarity Extraction,” *IEEE*

Access, vol. 12, pp. 10586–10607, 2024, doi: 10.1109/ACCESS.2023.3322103.

[193] A. Bans-Akutey and B. M. Tiimub, “Triangulation in Research,” *Acad. Lett.*, Aug. 2021, doi: 10.20935/AL3392.

APPENDICES

Appendix 1: Sample Experimental Checklist

Preprocessing of datasets done prior to running the experiments:

Amazon mobile phone reviews datasets preprocessing done includes the following:

i. Tokenization, stemming, lemmatization, handling special characters and data splitting
ii. Validation on the preprocessed data to ensure semantic consistency, retention of domain specific terms and idiomatic expressions
iii. Feature extraction transforms the textual data into numerical representations.
iv. Feature selection optimizes the feature space

Number of runs: **10**

Initialization techniques used across all the sentiment analysis models: **feature extraction, feature selection and data augmentation**

State-of-the-art sentiment analysis model	**Technique used**			**Evaluation Metrics**			**Check list**
	Feature Extraction	Feature Selection	data augmentation	Accuracy	Cohens Kappa	Mathews correlation	Done / Not done
LSTM Model	FastText	RFE	Back translation	0.84	0.82	0.78	✓
Hybrid (LSTM and BERT)	FastText	TF-IDF	Back translation	0.92	0.89	0.90	✓
Hybrid(LSTM and Naïve Nayes)	Word2vec	RFE	Synonym replacement	0.88	0.86	0.79	✓
Support vector Machine	Word2vec	TF-IDF	Synonym replacement	0.76	0.78	0.81	✓

Appendix 2: Sample Experiments

Sentiment Analysis Model	Feature Extraction Method	Feature Selection Method	Data Augmentation Strategy
SVM	FastText	RFE	Back translation
SVM	FastText	RFE	Synonym replacement
SVM	FastText	TF-IDF	Back translation
SVM	FastText	TF-IDF	Synonym replacement
SVM	Word2Vec	RFE	Back translation
SVM	Word2Vec	RFE	Synonym replacement
Naïve Bayes	FastText	RFE	Back translation
Naïve Bayes	FastText	RFE	Synonym replacement
Naïve Bayes	FastText	TF-IDF	Back translation
Naïve Bayes	Word2Vec	RFE	Synonym replacement
Naïve Bayes	Word2Vec	TF-IDF	Back translation
Naïve Bayes	Word2Vec	TF-IDF	Synonym replacement
BERT	FastText	RFE	Back translation
BERT	FastText	RFE	Synonym replacement
BERT	FastText	TF-IDF	Back translation
BERT	Word2Vec	RFE	Synonym replacement
BERT	Word2Vec	TF-IDF	Back translation
BERT	Word2Vec	TF-IDF	Synonym replacement
LSTM	FastText	RFE	Back translation
LSTM	FastText	RFE	Synonym replacement
LSTM	FastText	TF-IDF	Back translation
LSTM	Word2Vec	RFE	Synonym replacement
LSTM	Word2Vec	TF-IDF	Back translation
LSTM	Word2Vec	TF-IDF	Synonym replacement
SVM + Naïve Bayes	FastText	RFE	Back translation
SVM + Naïve Bayes	FastText	RFE	Synonym replacement
SVM + Naïve Bayes	FastText	TF-IDF	Back translation
SVM + Naïve Bayes	FastText	TF-IDF	Synonym replacement
SVM + Naïve Bayes	Word2Vec	RFE	Back translation
SVM + Naïve	Word2Vec	RFE	Synonym replacement

Bayes			
SVM + Naïve Bayes	Word2Vec	TF-IDF	Back translation
BERT + LSTM	FastText	RFE	Synonym replacement
BERT + LSTM	FastText	TF-IDF	Back translation
BERT + LSTM	Word2Vec	RFE	Synonym replacement
BERT + LSTM	Word2Vec	TF-IDF	Back translation
BERT + LSTM	Word2Vec	TF-IDF	Synonym replacement
SVM + BERT	FastText	RFE	Back translation
SVM + BERT	FastText	RFE	Synonym replacement
SVM + BERT	FastText	TF-IDF	Synonym replacement
SVM + BERT	Word2Vec	RFE	Back translation
SVM + BERT	Word2Vec	RFE	Synonym replacement
SVM + LSTM	Word2Vec	RFE	Back translation
SVM + LSTM	Word2Vec	RFE	Synonym replacement
SVM + LSTM	Word2Vec	TF-IDF	Back translation
SVM + LSTM	Word2Vec	TF-IDF	Synonym replacement
Naïve Bayes + BERT	FastText	RFE	Back translation
Naïve Bayes + BERT	FastText	RFE	Synonym replacement
Naïve Bayes + BERT	FastText	TF-IDF	Back translation
Naïve Bayes + BERT	Word2Vec	RFE	Synonym replacement
Naïve Bayes + BERT	Word2Vec	TF-IDF	Synonym replacement
Naïve Bayes + LSTM	FastText	RFE	Back translation
Naïve Bayes + LSTM	FastText	TF-IDF	Back translation
Naïve Bayes + LSTM	Word2Vec	RFE	Synonym replacement
Naïve Bayes + LSTM	Word2Vec	TF-IDF	Back translation
Naïve Bayes + LSTM	Word2Vec	TF-IDF	Synonym replacement

Appendix 3: Code Segments

```
# Import necessary libraries
!pip install transformers
!pip install tensorflow
!pip install scikit-learn
# Import necessary libraries
import numpy as np
import pandas as pd
import tensorflow as tf
import seaborn as sns
import matplotlib.pyplot as plt
from transformers import BertTokenizer, TFBertModel
from tensorflow.keras.layers import LSTM, Dense, Dropout
from tensorflow.keras.models import Model
from sklearn.metrics import matthews_corrcoef, cohen_kappa_score, mean_squared_error, accuracy_score
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
from tensorflow.keras.preprocessing.sequence import pad_sequences
rom google.colab import files
uploaded = files.upload()
# Load Amazon Reviews dataset
data = pd.read_csv('AmazonPhoneReviews.csv')
# Combine 'title' and 'body' columns to form the review text
data['review_text'] = data['title'] + " " + data['body']

# Generate sentiment labels based on the 'rating' column
# Example: Positive (1 for rating ≥ 4), Negative (0 for rating ≤ 2)
data['label'] = data['rating'].apply(lambda x: 1 if x >= 4 else (0 if x <= 2 else None))

# Filter out neutral ratings
data = data.dropna(subset=['label'])

# Extract the reviews and labels for sentiment analysis
reviews = data['review_text']
labels = data['label']
# Tokenizer and model preparation
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
bert_model = TFBertModel.from_pretrained('bert-base-uncased')
# Ensure all reviews are strings and remove invalid entries
reviews_cleaned = reviews.dropna().astype(str).tolist()

# Preprocessing: Tokenization and padding
def preprocess(texts):
    encodings = tokenizer(texts, truncation=True, padding=True, max_length=128, return_tensors="tf")
    return encodings['input_ids'], encodings['attention_mask']

# Pass the cleaned list of reviews to the function
input_ids, attention_masks = preprocess(reviews_cleaned)

# Data augmentation: Contextual back translation with Seq2Seq perturbations
from transformers import MarianMTModel, MarianTokenizer

# Define source and target languages for back translation
source_lang = 'en'  # Original language
```

```
target_lang = 'fr'  # Intermediate language

# Load MarianMT models for translation
model_name_src_to_tgt = f'Helsinki-NLP/opus-mt-{source_lang}-{target_lang}'
model_name_tgt_to_src = f'Helsinki-NLP/opus-mt-{target_lang}-{source_lang}'

tokenizer_src_to_tgt = MarianTokenizer.from_pretrained(model_name_src_to_tgt)
model_src_to_tgt = MarianMTModel.from_pretrained(model_name_src_to_tgt)

tokenizer_tgt_to_src = MarianTokenizer.from_pretrained(model_name_tgt_to_src)
model_tgt_to_src = MarianMTModel.from_pretrained(model_name_tgt_to_src)

# Function for back translation with truncation
def back_translate(text, model_src_to_tgt, model_tgt_to_src, tokenizer_src_to_tgt, tokenizer_tgt_to_src, max_length=512):
    # Translate from source to target language with truncation
    tokens = tokenizer_src_to_tgt(text, return_tensors="pt", truncation=True, max_length=max_length, padding=True)
    translation = model_src_to_tgt.generate(**tokens)
    tgt_text = tokenizer_src_to_tgt.batch_decode(translation, skip_special_tokens=True)[0]

    # Translate back from target to source language with truncation
    tokens = tokenizer_tgt_to_src(tgt_text, return_tensors="pt", truncation=True, max_length=max_length, padding=True)
    back_translation = model_tgt_to_src.generate(**tokens)
    augmented_text = tokenizer_tgt_to_src.batch_decode(back_translation, skip_special_tokens=True)[0]

    return augmented_text

# Split long reviews into chunks for better handling
def split_text_into_chunks(text, tokenizer, max_length=512):
    tokens = tokenizer(text, return_tensors="pt", truncation=False, padding=False)['input_ids'][0]
    chunks = [tokens[i:i + max_length] for i in range(0, len(tokens), max_length)]
    return [tokenizer.decode(chunk, skip_special_tokens=True) for chunk in chunks]

# Process reviews with back translation
augmented_reviews = []
for review in reviews:  # Assuming `reviews` is a list of strings
    if len(tokenizer_src_to_tgt(review)['input_ids']) > 512:
        # Split long reviews into chunks
        chunks = split_text_into_chunks(review, tokenizer_src_to_tgt)
        back_translated_chunks = [back_translate(chunk, model_src_to_tgt, model_tgt_to_src, tokenizer_src_to_tgt, tokenizer_tgt_to_src) for chunk in chunks]
        augmented_review = " ".join(back_translated_chunks)
    else:
        # Direct back translation for short reviews
        augmented_review = back_translate(review, model_src_to_tgt, model_tgt_to_src, tokenizer_src_to_tgt, tokenizer_tgt_to_src)
    augmented_reviews.append(augmented_review)

# Display a few examples of original vs. augmented reviews
for original, augmented in zip(reviews[:3], augmented_reviews[:3]):
    print(f'Original: {original}')
    print(f'Augmented: {augmented}\n')

# Define the hybrid BERT + LSTM model after back translation using Seq2seq pertubations
```

```
class BERT_LSTM_Model(Model):
    def __init__(self):
        super(BERT_LSTM_Model, self).__init__()
        self.bert = bert_model
        self.lstm = LSTM(64, return_sequences=False)
        self.dropout = Dropout(0.3)
        self.dense = Dense(1, activation='sigmoid')

    def call(self, inputs):
        input_ids, attention_mask = inputs
        outputs = self.bert(input_ids, attention_mask=attention_mask)
        x = self.lstm(outputs.last_hidden_state)
        x = self.dropout(x)
        x = self.dense(x)
        return x
# Initialize and compile the Hybrid Novel model
model = BERT_LSTM_Model()
model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=2e-5),
loss='binary_crossentropy', metrics=['accuracy'])

# Train the model
history = model.fit([input_ids, attention_masks], labels, epochs=3, batch_size=16,
validation_split=0.2)
# Predict and evaluate
predictions = (model.predict([input_ids, attention_masks]) > 0.5).astype(int)
# Quantitative Evaluation
mcc = matthews_corrcoef(labels, predictions)
kappa = cohen_kappa_score(labels, predictions)
mse = mean_squared_error(labels, predictions)
accuracy = accuracy_score(labels, predictions)

print(f"Matthews Correlation Coefficient: {mcc}")
print(f"Cohen's Kappa: {kappa}")
print(f"MSE: {mse}")
print(f"Accuracy: {accuracy}")
```

Appendix 4: Directorate of Postgraduate Studies Research Approval Letter

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:** 7th November, 2023

Ms. Millicent Kathambi Murithi (SC500/5532/2020)
School of Computing and Information Technology,
Murang'a University of Technology,
P.O. Box 75-10,200
Murang'a.

Dear Ms. Murithi,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 13th October, 2023 considered and approved your Doctoral research proposal entitled ***"Deep Naïve Bayes Classifier for Handling the Polarity Shift Problem in Sentiment Analysis of the Mobile Phone Brands' Reviews"*** and appointed the following as supervisors:

1. Dr. Aaron Mogere Oirere - Murang'a University of Technology
2. Dr. Rachael Njeri Ndung'u - 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;
iii. Ensure quality work all through;

iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.
v. Publish two 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: Research Permit from NACOSTI



REPUBLIC OF KENYA

NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **790278**

Date of Issue: **11/July/2024**

RESEARCH LICENSE

This is to Certify that Ms.. MILLICENT KATHAMBI MURITHI 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 Baringo, Bomet, Bungoma, Busia, Elgeyo-Marakwet, Embu, Garissa, Homabay, Isiolo, Kajiado, Kakamega, Kericho, Kiambu, Kilifi, Kirinyaga, Kisii, Kisumu, Kitui, Kwale, Laikipia, Lamu, Machakos, Makueni, Mandera, Marsabit, Meru, Migori, Mombasa, Muranga, Nairobi, Nakuru, Nandi, Narok, Nyamira, Nyandarua, Nyeri, Samburu, Siaya, Taita-Taveta, Tanariver, Tharaka-Nithi, Transnzoia, Turkana, Uasin-Gishu, Vihiga, Wajir, Westpokot on the topic: DEEP NAÏVE BAYES CLASSIFIER FOR HANDLING THE POLARITY SHIFT PROBLEM IN SENTIMENT ANALYSIS OF THE MOBILE PHONE BRANDS' REVIEWS for the period ending : 11/July/2025.

License No: **NACOSTI/P/24/37594**

790278

Applicant Identification Number

Director General
NATIONAL COMMISSION FOR SCIENCE,TECHNOLOGY & INNOVATION

Verification QR Code

Appendix 6: Publications

The following papers have been published from this thesis:

[1] M. K. Murithi, A. M. Oirere, R. N. Ndung'u , "A Systematic Review of the Sentiment Analysis Models Used in Handling Polarity Shift". International Journal of Computing Sciences Research, 8, 2635-2676, 2024.

[2] M. K. Murithi, A. M. Oirere &, R. N. Ndung'u. "Empirical Analysis of the State-of-the-Art Models for Handling Polarity Shifts Due to Implicit Negation in Mobile Phone Reviews". International Journal of Computing Sciences Research, 9, 3392-3417, 2025.

[3] M. K. Murithi, A. M. Oirere, R. N. Ndung'u, Murithi, "Hybrid Deep Learning Model for Detecting Polarity Shift due to Implicit Negation in Sentiment Analysis of Mobile Phone Brands' Reviews". Computational Linguistics Journal, Manuscript under review, 2025.