

TABLE OF CONTENTS

APPROVAL PAGE	i
PLAIGARISM STATEMENT	ii
ABSTRACT	iii
INTISARI	iv
TABLE OF CONTENTS	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Scope	3
1.4 Research Objectives	4
1.5 Research Benefits	4
1.6 Research Methodology	4
CHAPTER II LITERATURE REVIEW	6
CHAPTER III THEORETICAL BASIS	15
3.1 Sentiment Analysis	15
3.1.1 IMDb Movie Reviews	16
3.2 Pre-Processing	16
3.2.1 Noise Removal	17
3.2.2 Case Folding	17
3.2.3 Contraction Expansion	17
3.2.4 Slang Expansion	18
3.2.5 Spell Checker	18
3.2.6 Tokenization	19
3.2.7 Stopword removal	19
3.2.8 Stemming	19
3.2.9 Lemmatization	20
3.3 Feature Extraction	20
3.3.1 Term Frequency-Inverse Document Frequency	21
3.3.2 Bag-of-Words	22
3.3.3 Feature Selection	22
3.4 Word Embedding	23
3.4.1 Word2Vec	24
3.4.2 Doc2Vec	24
3.4.3 BERT Embedding	25
3.5 Machine Learning	26
3.5.1 Decision Tree	26
3.5.2 Logistic Regression	26
3.5.3 Naïve Bayes	27
3.5.4 Support Vector Machines	28
3.6 Deep Learning	28
3.6.1 Recurrent Neural Network	29
3.6.2 Long Short-Term Memory	29
3.6.3 Bidirectional Long Short-Term Memory	31

3.6.4 Bidirectional Encoder Representations from Transformers	32
3.7 Attention Mechanisms.....	33
3.7.1 Additive Attention	33
3.7.2 Scaled Dot-Product Attention.....	33
3.8 Hybrid Models.....	34
3.8.1 Word2Vec-LSTM	34
3.8.2 Doc2Vec-LSTM.....	35
3.8.3 Word2Vec-BiLSTM with Attention Mechanism.....	35
3.9 Evaluation Metrics	36
3.9.1 Accuracy	37
3.9.2 Precision.....	38
3.9.3 Recall	38
3.9.4 F1-Score.....	39
3.9.5 K-fold Cross-validation	39
CHAPTER IV METHODOLOGY.....	40
4.1 Research Description.....	40
4.2 Problem Analysis	40
4.3 Data Acquisition.....	40
4.4 Model Design	41
4.4.1 Experimental Setup	41
4.4.2 Algorithm Overview	42
4.4.3 Pre-processing.....	43
4.4.4 Data Splitting	44
4.4.5 K-fold Cross-validation	44
4.4.6 TF-IDF+SVM Model Design	45
4.4.7 Word2Vec-BiLSTM Model Design	46
4.4.8 Word2Vec-BiLSTM-Att Model Design.....	47
4.5 Model Analysis and Performance Testing	48
CHAPTER V RESEARCH IMPLEMENTATION.....	50
5.1 Importing Libraries	50
5.2 Pre-processing	53
5.2.1 Initial Data Cleaning and Noise Removal	53
5.2.2 Text Normalization	54
5.2.3 Spell Checking.....	55
5.2.4 Tokenization, Stopword Removal, Lemmatization	56
5.2.5 Full Pre-processing Implementation.....	56
5.2.6 Secondary Data Cleaning.....	57
5.3 Dataset Splitting	57
5.4 Feature Extraction	58
5.4.1 Term Frequency Inverse Document Frequency (TF-IDF).....	58
5.4.2 Word2Vec	58
5.5 TF-IDF+SVM Model	59
5.5.1 TF-IDF+SVM Model Architecture.....	59
5.5.2 TF-IDF+SVM Hyperparameter Tuning	59
5.5.3 TF-IDF+SVM Model Training with K-fold Cross-validation.....	63
5.6 Word2Vec-BiLSTM Model	64

5.6.1 Word2Vec-BiLSTM Model Architecture.....	64
5.6.2 Word2Vec-BiLSTM Model Fine-Tuning Configurations.....	65
5.6.3 Word2Vec-BiLSTM Model Training with K-fold Cross-validation.....	67
5.7 Word2Vec-BiLSTM with Attention Mechanisms.....	69
5.7.1 Word2Vec-BiLSTM with Additive Attention Layer.....	70
5.7.2 Word2Vec-BiLSTM with Scaled Dot-Product Attention Layer	71
5.7.3 Word2Vec-BiLSTM with Attention on Underperforming Fold.....	73
5.8 Performance Metrics	73
5.8.1 Accuracy	73
5.8.2 F1-score	73
5.8.3 Training Time	74
CHAPTER VI RESULTS AND DISCUSSION	75
6.1 Data Acquisition.....	75
6.2 Pre-processing	75
6.2.1 Initial Data Cleaning and Noise Removal	75
6.2.2 Text Normalization Results	76
6.2.3 Spell Checker Results	77
6.2.4 Pre-processing Pipeline Results.....	78
6.2.5 Secondary Data Cleaning.....	78
6.3 Data Splitting Results.....	79
6.4 Feature Extraction Results	79
6.4.1 TF-IDF Results	79
6.4.2 Word2Vec Results	80
6.5 TF-IDF+SVM Model Results	81
6.5.1 Initial Hyperparameter Tuning Testing Results.....	81
6.5.2 TF-IDF+SVM Model Hyperparameter Tuning Results	82
6.5.3 Final TF-IDF+SVM Model Performance	82
6.6 Word2Vec-BiLSTM Configurations Comparison.....	82
6.7 Word2Vec-BiLSTM Model Performance.....	85
6.8 Word2Vec-BiLSTM with Attention Mechanism Model	87
6.8.1 Additive Attention Mechanism.....	87
6.8.2 Scaled Dot Product Attention Mechanism.....	89
6.8.3 Attention Mechanisms on Underperforming Fold.....	90
6.9 Comparative Analysis of All Models.....	92
6.10 Discussion	93
6.11 Research Limitations.....	93
CHAPTER VII CONCLUSION	95
7.1 Conclusion.....	95
7.2 Future Work	95
REFERENCES.....	97
APPENDIX	103

LIST OF TABLES

Table 2.1 Overview of Sentiment Analysis Methods in Previous Studies.....	13
Table 5.1 Preliminary Hyperparameter Testing on 5,000 Sample Subset	60
Table 5.2 Final Hyperparameter Tuning Grid for SVM.....	62
Table 5.3 Non-Stacked Word2Vec-BiLSTM Configurations.....	67
Table 5.4 Stacked Word2Vec-BiLSTM Configurations.....	67
Table 6.1 Dataset Split	79
Table 6.2 Linear Kernel – Initial Hyperparameter Tuning Testing (5,000 Subset)	81
Table 6.3 RBF Kernel – Initial Hyperparameter Tuning Testing (5,000 Subset)	81
Table 6.4 GridSearchCV Results for TF-IDF+SVM Model (10,000 Sample)	82
Table 6.5 Best Performing Epoch per Word2Vec-BiLSTM Configuration	83
Table 6.6 Performance of Word2Vec-BiLSTM model across 3-fold cross- validation.....	86
Table 6.7 Performance of Word2Vec-BiLSTM model with additive attention layer across 3-fold cross-validation	88
Table 6.8 Performance of Word2Vec-BiLSTM model with scaled dot-product attention layer across 3-fold cross-validation	89
Table 6.9 Performance of attention mechanisms on underperforming data splits	91
Table 6.10 Model Results.....	92

LIST OF FIGURES

Figure 3.1 Stemming process (Başa and Başarslan, 2023)	20
Figure 3.2 Lemmatization process (Dahir and Alkindy, 2023).....	20
Figure 3.3 Illustration of Bag-of-Word process (Dahir and Alkindy, 2023)	22
Figure 3.4 CBoW and skip-gram Word2Vec architecture (Bilgin and Şentürk, 2017)	24
Figure 3.5 DM and DBoW model architecture (Bilgin and Şentürk, 2017)	25
Figure 3.6 Visual representation of SVM (García-Gonzalo <i>et al.</i> , 2016)	28
Figure 3.7 Cell structure of LSTM unit (Cai <i>et al.</i> , 2024)	30
Figure 3.8 BiLSTM architecture (Li <i>et al.</i> , 2020).....	31
Figure 4.1 Sample of IMDb movie review dataset	41
Figure 4.2 Flowchart of the research method.....	42
Figure 4.3 Pre-processing flow used in this research.....	44
Figure 5.1 List of all libraries used in the research	52
Figure 5.2 Code implementation for data cleaning and noise removal.....	53
Figure 5.3 Code implementation for text normalization	55
Figure 5.4 Code implementation for spell-checking	55
Figure 5.5 Code implementation for tokenization, stopword removal, and lemmatization.....	56
Figure 5.6 Code implementation for main pre-processing functions.....	57
Figure 5.7 Code implementation for TF-IDF implementation.....	58
Figure 5.8 Code implementation for custom Word2Vec model	59
Figure 5.9 Code implementation for base TF-IDF + SVM model.....	59
Figure 5.10 Code implementation for initial hyperparameter tuning.....	61
Figure 5.11 Code implementation to perform hyperparameter tuning for SVM .	62
Figure 5.12 Full code implementation for TF-IDF+SVM model training.....	64
Figure 5.13 Word2Vec-BiLSTM model architecture	68
Figure 5.14 Code implementation for training and evaluation of Word2Vec-BiLSTM model with 3-fold cross-validation.....	69
Figure 5.15 Code implementation for the additive attention mechanism used	71
Figure 5.16 Code implementation for scaled dot-product attention mechanism used	72
Figure 6.1 Sample of review before and after noise removal	76
Figure 6.2 Sample review before and after text normalization (a) Review contains contractions (b) Review contains slang and contractions	77
Figure 6.3 Sample review before and after applying spell checking	77
Figure 6.4 Sample review before and after pre-processing.....	78
Figure 6.5 Reviews that became identical after pre-processing.....	79
Figure 6.6 Top unigrams and bigrams by mean TF-IDF score.....	80
Figure 6.7 Loss and accuracy curves for configurations 1 to 7, labeled as subfigures (a) to (g), respectively	85
Figure 6.8 Loss and accuracy curves for Word2Vec-BiLSTM model using 3-fold cross-validation, labeled as subfigures (a), (b), (c) for each fold, respectively	87
Figure 6.9 Loss and accuracy curves for Word2Vec-BiLSTM model with additive mechanism using 3-fold cross-validation, labeled as subfigures (a), (b), (c) for each fold, respectively.....	88



Figure 6.10 Loss and accuracy curves for Word2Vec-BiLSTM model with scaled dot-product attention using 3-fold cross-validation, labeled as subfigures (a), (b), (c) for each fold, respectively.....	90
Figure 6.11 Loss and accuracy curves for the Word2Vec-BiLSTM model with additive attention on fold 3	91
Figure 6.12 Loss and accuracy curves for the Word2Vec-BiLSTM model with scaled dot-product attention on fold 3.....	92