

TABLE OF CONTENTS

THESIS APPROVAL SHEET.....	2
TABLE OF CONTENTS.....	3
LIST OF TABLES.....	6
LIST OF FIGURES	7
LIST OF ALGORITHMS & CODINGS	10
ABSTRACT	11
PLAGIARISM STATEMENT	12
FOREWORD.....	13
CHAPTER I INTRODUCTION	16
1.1 Background	16
1.2 Research Problem.....	18
1.3 Research Scope	18
1.4 Research Objectives	19
1.5 Research Benefits.....	20
CHAPTER II LITERATURE REVIEW	22
CHAPTER III THEORETICAL FOUNDATION.....	28
3.1 Theoretical Basis of Machine Learning Models.....	28
3.2 Sentiment Classification Models	28
3.2.1 Lexicon-based Method (VADER).....	29
3.2.2 Supervised Machine Learning-based Method	31
3.2.3 Hybrid Sentiment Classification Model	31
3.3 XGBoost.....	32
3.4 Relevance of Hybrid Models in Social Media Sentiment Analysis..	33
3.5 Data Collection.....	34
3.6 Data Preprocessing	34
3.7 Sentiment Analysis Implementation	35
3.8 Model Development.....	36
3.8.1 Feature Engineering.....	36
3.9 Model Training and Testing.....	38
3.10 Model Validation.....	39
3.10.1 Cross-Validation	39
3.10.2 Baseline Comparisons	40
3.11 Model Optimization.....	40

3.11.1	Hyperparameter Tuning	40
3.11.2	Critical Hyperparameters	41
3.12	Model Evaluation.....	43
3.12.1	Accuracy.....	43
3.12.2	Precision	43
3.12.3	Recall (Sensitivity or True positive rate).....	44
3.12.4	F1-Score	44
3.12.5	Confusion Matrix	44
3.12.6	ROC Curves.....	45
3.12.7	AUC Scores	46
CHAPTER IV	RESEARCH METHODOLOGY	47
4.1	Data Collection.....	50
4.1.1	Data Source.....	50
4.1.2	Data Collection Period (Timeframe)	50
4.2	Data Preprocessing.....	51
4.2.1	Data Cleaning	51
4.2.2	Language Normalization	52
4.2.3	Tokenization and Text Processing (for TF-IDF)	53
4.2.4	Stop Words Removal and Lemmatization (for TF-IDF)	53
4.3	Dual Feature Extraction	54
4.3.1	VADER (SCompound Score, Spos, Sneu, Sneg)	55
4.3.2	TF-IDF (Matrix Score).....	57
4.4	Model Development (XGBoost).....	58
4.4.1	Model Architecture.....	58
4.4.2	Hyperparameter Tuning.....	66
4.4.3	Validation	66
4.5	Model Evaluation.....	67
4.6	Ethical Considerations	67
CHAPTER V	IMPLEMENTATION.....	68
5.1	Data Collection.....	68
5.2	Data Pre-processing.....	74
5.2.1	cleaning().....	78
5.2.2	normalization()	80
5.2.3	tokenize_and_filter().....	83
5.2.4	lemmatization().....	86

5.2.5	preprocessed_vader()	89
5.2.6	preprocessed_tfidf()	90
5.2.7	process_tweet_dataset()	92
5.2.8	analyze_preprocessing_results()	93
5.3	Feature Extraction	97
5.3.1	Labelling (VADER)	97
5.3.2	Manual Label Adjustment and Final Distribution	115
5.3.3	Feature Preparation for Model Training	117
5.3.4	Feature Engineering (Numerical Features)	120
5.3.5	TF-IDF Vectorization	123
5.3.6	Hybrid Feature Concatenation	126
5.4	Model Development (XGBoost)	128
5.4.1	Label Encoding into Numerical Values	128
5.4.2	Splitting Dataset	129
5.4.3	Hyperparameter-Tuning XGBoost	130
5.5	Model Evaluation	135
5.5.1	Extract Best Model and Generate Predictions	135
5.5.2	Accuracy Assessment	135
5.5.3	Cross-Validation Assessment	136
5.5.4	Classification Report	136
5.5.5	Most Important Features	136
5.5.6	Data Visualization	138
CHAPTER VI DISCUSSIONS & RESULTS		143
6.1	Model Performance and Validation	143
6.2	Performance Improvement Analysis	145
6.3	Classification Report	145
6.4	Feature Importance	147
6.5	Data Visualization	151
6.2.1	Confusion Matrix	151
6.2.2	Wordcloud (TF-IDF words)	152
6.2.3	ROC Curves and AUC Scores	153
CHAPTER VII CONCLUSION & SUGGESTIONS		155
7.1	Conclusion	155
7.2	Suggestions and Future Works	158
REFERENCES		159