

TABLE OF CONTENTS

THESIS COMPARATIVE STUDY OF DATA PREPROCESSING METHODS FOR CODE-MIXED SENTIMENT ANALYSIS IN BAHASA INDONESIA	i
ENDORSEMENT PAGE	ii
STATEMENT OF ORIGINALITY	iii
PREFACE	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	vii
LIST OF TABLES	viii
ABSTRACT	ix
INTISARI	x
CHAPTER I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	4
1.3 Research Objectives	4
1.4 Research Scope and Limitation	5
1.5 Research Benefits	5
CHAPTER II LITERATURE REVIEW	6
CHAPTER III THEORETICAL FRAMEWORK	14
3.1 Sentiment Analysis	14
3.1.1 Data Code-Mixed Indonesia-English	15
3.2 Data Preprocessing	15
3.2.1 Language Identification	16
3.2.2 Lexical Normalization	17
3.2.3 Translation	18
3.2.4 Tokenization	19
3.2.5 Stop Word Removal	19
3.3 SVM (Support Vector Machine)	20
3.4 RF (Random Forest)	21
3.5 TF-IDF Vectorizer	22
3.6 LSTM (Long Short-Term Memory)	22
3.7 One-hot Encoding Vectorization	23
3.8 Performance Based Evaluation	23
3.8.1 Label-Based Evaluation	24
3.8.2 Accuracy	24
3.8.3 Precision	24
3.8.4 Recall	25
3.8.5 F1 Score	25

CHAPTER IV RESEARCH METHODOLOGY	26
4.1 Research Description	26
4.2 Research Steps	27
4.2.1 Dataset	29
4.2.2 Data Cleaning Preprocessing	30
4.2.3 Preprocessing Methods Scenarios	35
4.2.4 Data Splitting and Random State	38
4.2.5 Model Training and Classification	39
4.2.6 Performance Evaluation	40
CHAPTER V	42
IMPLEMENTATION	42
5.1 Dataset	43
5.2 Data Cleaning	45
5.3 Data Preprocessing – Defining the Scenarios	47
5.3.1 Lexical Normalization	47
5.3.3 Identify Language	49
5.3.4 Stop Word Removal	50
5.3.5 Translation	51
5.3.5 Data Splitting for All Models	51
5.4 SVM Model	51
5.6 RF Model	55
5.6 LSTM Model	57
5.7 Defining Each Scenario	60
CHAPTER VI RESULT AND DISCUSSION	64
6.1 Dataset Category	64
6.2 Splitting of Dataset	66
6.3 Performance Measure Result for Each Scenario	67
6.4 Analysis Summary	75
CHAPTER VII	77
CONCLUSION AND SUGGESTIONS	77
7.1 Conclusion	77
7.2 Suggestions for Future Work:	77
BIBLIOGRAPHY	79
ATTACHMENTS	83

LIST OF FIGURES

Figure 1.1: Number of Internet Users in Indonesia in Millions (Asosiasi Penyelenggara Jasa Internet Indonesia, 2023)	1
Figure 1.2 Interpretation of Types of Data Preprocessing (Alasy & Bhaya, 2017)	4
Figure 2.1: Proposed Pipeline for Series of Normalization Varma and Mamidi (2021). 10	
Figure 3.1 Common Pipeline for Sentiment Analysis (Varma & Mamidi, 2021)	14
Figure 3.2 Algorithm for Language Identification	17
Figure 3.3 Algorithm for Lexical Normalization	17
Figure 3.4 Algorithm for Translation	18
Figure 3.5 Algorithm for Tokenization	19
Figure 3.5 Algorithm for Stop Word Removal	20
Figure 3.6 Diagram of Random Forest Decision Tree	21
Figure 4.1 Research Flow Diagram	26
Figure 4.2 General Research Steps with Scenarios	29
Figure 4.3 Data Cleaning Preprocessing Step	31
Figure 4.4 Example of Confusion Matrix Comparison (Astuti <i>et al</i> , 2023)	40
Figure 5.1 Source Code for Uploading and Align Parameters of Dataset	44
Figure 5.1 (Cont) Source Code for Uploading and Align Parameters of Dataset	45
Figure 5.2 Source Code for Data Preprocessing 1 (Data Cleaning)	46
Figure 5.3.1 Indonesian Lexical Normalization	47
Figure 5.3.2 Source Code of Normalization Function	48
Figure 5.3.3 Source Code of Identify Language Function	49
Figure 5.3.4 Source Code of Stop Word Removal Function	50
Figure 5.3.5 Example of Translation Using Google Sheets	51
Figure 5.3.6 Consistent Data Splitting with Random State	51
Figure 5.4.1 Source Code TF-IDF and One-hot Encoding Vectorization for SVM	53
Figure 5.4.2 Source Code for SVM Model Training and Classification	53
Figure 5.4.3 Source Code for SVM Model Classification Report	54
Figure 5.6.1 TF-IDF and One Hot Encoder Vectorizer for RF Model	55
Figure 5.6.2 Training and Classification for RF Model	56
Figure 5.6.1 Source Code for LSTM Vectorization with One-hot Encoding	58
Figure 5.6.2 Source Code for LSTM Training and Classification	59
Figure 5.6.3 Source Code for LSTM Classification Report	59
Figure 6.1 Sentiment Label Distribution Graph	65
Figure 6.2 Bar Graph of Train Set Label Distribution	66
Figure 6.3 Bar Graph of Test Set Label Distribution	67

LIST OF TABLES

Table 2.1: Literature Review	12
Table 4.1 Scenario Process for Pre-Processing	28
Table 4.2 Examples of Sentiment (Astuti <i>et al</i> , 2023)	30
Table 4.3 Examples of Context	30
Table 4.4 Tweet with Lowercasing Step	32
Table 4.5 Tweet with Retweet Removal Step	32
Table 4.6 Tweet with Mention Removal Step	33
Table 4.7 Tweet with Mention Removal Step	33
Table 4.8 Tweet with URL Removal Step	34
Table 4.10 Tweet with Non-Alpha Numeric Removal Step	34
Table 4.11 Tweet with Tokenization Step	35
Table 4.13 Tweet with Lexical Normalization	36
Table 4.12 Tweet with Language Identification	36
Table 4.14 Some of Stop Words from the Library	37
Table 4.15 Tweet with Stop Word Removal	37
Table 4.16 Tweet with Translation	38
Table 4.18 Matrix Output for Positive Label / Class (From Figure 4.4)	41
Table 4.19 Calculation for the Performance Evaluation Matrix	41
Table 5. 1 Result of Annotation and Labeling, Astuti (2023)	43
Table 5. 2 Majority Voting Result, Astuti (2023)	44
Table 6.1 Dataset Sentiment Category	64
Table 6.4 Performance Result for SVM	68
Table 6.5 Performance Result for RF	71
Table 6.6 Performance Result for LSTM Model	73