

## CONTENTS

APPROVAL PAGE .....	ii
PLAGIARISM STATEMENT .....	iii
PREFACE .....	iv
CONTENTS.....	v
LIST OF TABLES .....	viii
LIST OF FIGURES .....	ix
LIST OF EQUATIONS .....	xi
<i>ABSTRACT</i> .....	xii
CHAPTER I INTRODUCTION .....	1
1.1 Background.....	1
1.2 Research Problem.....	4
1.3 Research Scope.....	4
1.4 Research Objective .....	5
1.5 Research Benefits .....	6
1.6 Writing Method .....	6
CHAPTER II LITERATURE REVIEW .....	8
CHAPTER III THEORETICAL BASIS .....	18
3.1 Natural Language Processing .....	18
3.2 Code-Mixed Text.....	18
3.2.1 Code-Mixing in Indonesian Social Media.....	19
3.3 Sentiment Analysis.....	19
3.3.1 Method Classification .....	19
3.3.1.1 Weak Supervision Auto-Labeling .....	20
3.3.2 Lexicon-Based.....	21
3.3.2.1 SentIL .....	21
3.3.3 Deep Learning.....	22
3.3.4 BiLSTM .....	22
3.3.5 Transformer.....	23
3.3.6 BERT.....	23
3.3.6.1 IndoBERT.....	24
3.3.6.2 IndoBERTweet.....	24
3.4 Data Preprocessing .....	25
3.4.1 Cleaning .....	25

3.4.1.1	Filtering .....	25
3.4.1.2	Lowercasing .....	25
3.4.1.3	Remove Noise .....	25
3.4.1.4	Punctuation Handling .....	26
3.4.1.5	Space Normalization .....	26
3.4.1.6	Number Conversion.....	26
3.4.2	Slang Normalization.....	26
3.4.3	Tokenization.....	27
3.5	Model Evaluation .....	27
3.5.1	Accuracy .....	27
3.5.2	Recall.....	28
3.5.3	Precision.....	28
3.5.4	F1-Score .....	28
3.5.5	Confusion Matrix .....	28
3.5.6	Cohen's Kappa .....	29
3.6	Gender Equality.....	31
CHAPTER IV RESEARCH METHODOLOGY .....		32
4.1	Literature Study .....	33
4.2	Research Setup .....	33
4.3	Research Design Overview .....	34
4.4	Dataset .....	35
4.5	Data Preprocessing .....	36
4.5.1	Filtering .....	37
4.5.2	Cleaning .....	37
4.5.3	Slang Normalization.....	38
4.5.4	Selective Translation.....	38
4.6	Lexicon Based (SentIL) Auto-Labeling .....	38
4.7	Data Preparation .....	39
4.8	Model Design Plan .....	39
4.8.1	BiLSTM .....	40
4.8.1	IndoBERT .....	41
4.8.2	IndoBERTweet.....	42
4.9	Model Implementation .....	43
4.10	Model Evaluation .....	44
4.10.1	Quantitative.....	44
4.10.2	Qualitative.....	44

4.11	Writing the Report .....	44
CHAPTER V IMPLEMENTATION .....		46
5.1	Data Collection .....	46
5.2	Data Preprocessing .....	52
5.2.1	Exploratory Data Analysis .....	52
5.2.2	Filtering .....	55
5.2.3	Text Cleaning .....	56
5.2.4	Slang Normalization.....	57
5.2.5	Selective Translation.....	58
5.3	Auto-Labeling with SenTIL .....	60
5.3.1	Manually Annotated Data .....	61
5.3.2	SenTIL Annotation Evaluation .....	62
5.4	Data Splitting.....	63
5.5	BiLSTM.....	65
5.6	IndoBERT.....	68
5.7	IndoBERTweet.....	71
5.8	Quantitative Evaluation .....	71
CHAPTER VI RESULTS AND DISCUSSION.....		72
6.1	Auto-Labeling with SenTIL .....	72
6.2	Data Splitting.....	74
6.3	Model Comparison .....	75
6.3.1	BiLSTM .....	76
6.3.2	IndoBERT .....	79
6.3.3	IndoBERTweet.....	83
6.4	Quantitative Evaluation .....	85
CHAPTER VII CONCLUSION AND FUTURE WORK .....		88
7.1	Conclusion.....	88
7.2	Future Work.....	89
REFERENCES .....		90