

TABLE OF CONTENT

ABSTRACT	i
INTISARI.....	ii
ACKNOWLEDGEMENT	iii
PLAGIARISM STATEMENT	iv
TABLE OF CONTENT	v
LIST OF TABLE	vii
LIST OF FIGURE.....	viii
CHAPTER I INTRODUCTION.....	1
1.1 Research Background	1
1.2 Research Problem.....	3
1.3 Research Objectives	3
1.4 Structure of the Research Report.....	4
1.5 Research Ethics	4
CHAPTER II LITERATURE REVIEW	6
CHAPTER III BASIC THEORY	10
3.1 Natural Language Processing (NLP) & Large Language Model (LLM).....	10
3.2 Sentiment Analysis & Text Classification	10
3.3 Text Representation.....	11
3.4 Traditional Text Classification Model.....	12
3.5 Logistic Regression	12
3.6 Random Forest	15
3.7 Naïve Bayes	17
3.8 K-Nearest Neighbour (KNN).....	20
3.9 Pre-trained Language Model (PLMs).....	23
3.10 Fine-Tuning.....	24
3.11 Transformer-Based Model.....	24
3.12 Bidirectional Encoder Representations from Transformers (BERT)	25
3.13 Bert Advancement	26
3.14 Evaluation Metrics	27
3.14.1 K-fold Cross-Validation	27
3.14.2 Confusion Matrix	28
3.14.3 Accuracy Score	29
3.14.4 Precision.....	29

3.14.5	Recall.....	30
3.14.6	F1-Score	30
CHAPTER IV RESEARCH METHODOLOGY		32
4.1	Research Description	32
4.2	Data Acquisition.....	32
4.2.1	Data Acquisition of Stocker Twitter Dataset	33
4.3	Data Pre-processing.....	35
4.4	Traditional Classification Models	39
4.4.1	Bag of Words Tokenisation	39
4.4.2	Logistic Regression Training	41
4.4.3	Random Forest Training	41
4.4.4	Naïve Bayes Training	42
4.4.5	K-Nearest Neighbours Training	42
4.4.6	K-Fold Cross-Validation	43
4.5	Pre-trained Language Models	43
4.5.1	BERT Tokenisation.....	43
4.5.2	BERT Fine-Tuning.....	44
CHAPTER V RESULT AND DISCUSSION		46
5.1	Logistic Regression Model	46
5.2	Random Forest Model	47
5.3	Naïve Bayes model.....	49
5.4	K-Nearest Neighbour Model.....	50
5.5	Fine-Tuned Pre-Trained Model.....	52
5.6	Models Analysis	53
CHAPTER VI CONCLUSION		57
6.1	Conclusion	57
6.2	Future Work	58
REFERENCE		60
APPENDIX		63

LIST OF TABLE

Table 4.1 Stocker Twitter Dataset	32
Table 4.2 PhraseBank Dataset	33
Table 4.3 Noise removal example table	35
Table 4.4 Table of data splitting	36
Table 4.5 Logistic regression parameters	40
Table 4.6 Random forest parameter	41
Table 4.7 Naïve Bayes parameters	41
Table 4.8 K-nearest neighbours parameters	41
Table 4.9 Fine-tuning parameters	44
Table 5.1 Logistic regression model accuracy	46
Table 5.2 Random Forest model accuracy	48
Table 5.3 Naïve Bayes model accuracy	49
Table 5.4 K-nearest neighbours model accuracy	51
Table 5.5 All model matrix	53

LIST OF FIGURE

Figure 3.1 Logistic Regression Algorithm	13
Figure 3.2 Decision Tree	14
Figure 3.3 Random Forest	14
Figure 3.4 Random Forest Algorithm	15
Figure 3.5 Naïve Bayes Algorithm	19
Figure 3.6 Illustration of the K-nearest neighbour model	20
Figure 3.7 K-Nearest Neighbour Algorithm	21
Figure 3.8 Model architecture of Transformer	24
Figure 3.9 Fine-tuning BERT architecture	25
Figure 3.10 Example of the matrix table	28
Figure 4.1 The flowchart of the research framework	32
Figure 4.2 Sample data of the Stocker Twitter Dataset	33
Figure 4.3 Sample data of the PhraseBank Dataset	34
Figure 4.4 Summary of Dataset Dataset	36
Figure 4.5 Positive sentiment wordcloud	37
Figure 4.6 Neutral sentiment wordcloud	37
Figure 4.7 Negative sentiment wordcloud	37
Figure 4.8 Bag of Words example	39
Figure 4.9 BERT tokenization example	43
Figure 5.1 Logistic regression confusion matrix	45
Figure 5.2 Random Forest confusion matrix	47
Figure 5.3 Naïve Bayes confusion matrix	48
Figure 5.4 K-nearest neighbours confusion matrix	50
Figure 5.5 Fine-tuning graph	52