

TABLE OF CONTENTS

UNDERGRADUATE THESIS	1
APPROVAL PAGE	3
PLAGIARISM STATEMENT	I
TABLE OF CONTENTS	IV
TABLE OF FIGURES	VIII
TABLE OF TABLES	IX
ABSTRACT	X
CHAPTER I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	3
1.3 Research Scope	3
1.4 Research Objectives	3
1.5 Research Benefits	4
CHAPTER II LITERATURE REVIEW	5
CHAPTER III THEORETICAL BASIS	11
3.1 Natural Language Processing	11

3.2 Sentiment Analysis	11
3.3 Pre-processing	12
3.3.1 Lower case	12
3.3.2 Stop word removal	13
3.3.3 Punctuation removal	13
3.3.4 Stemming	14
3.3.5 Bag Of Word	14
3.4 Classifier	14
3.4.1 Naive Bayes	14
3.4.2 Multinomial Naive Bayes	15
3.5 Semi-Supervised Learning	16
3.5.1 Seeding Theory	17
3.5.2 Seeding Clustering	17
3.5.3 Seeding Classification	17
3.5.4 Pseudo labelling	18
3.6 Performance Evaluation	19
3.6.1 Confusion Matrix	20
3.6.2 Accuracy	20
3.6.3 Precision	20
3.6.4 Recall	21
3.4.5 F-1 Score	21
CHAPTER IV RESEARCH METHODOLOGY	22
4.1 Research Description	22
4.2 Tools and Material	23
4.2.1 Tools	23
4.2.2 Materials	23

4.3 Research Phase	24
4.4 System Design	28
4.4.1 Data	28
4.4.2 Pre-processing	30
4.4.3 Supervised Multinomial Naive Bayes Model	31
4.4.4 Semi Supervised implementing Seed theory	32
4.4.5 Additional Experiment	35
4.4.6 Evaluation	36
CHAPTER V IMPLEMENTATION	37
5.1 Dataset	37
5.2 Data Pre-processed	38
5.2.1 Data Cleaning	38
5.2.2 Count Vectorizer	39
5.3 Machine Learning (1st model)	40
5.4 Seeding	41
5.4.1 All prediction from training data. (2 nd Model)	42
5.4.2 Prediction based on deviant score (3 rd Model)	43
5.4.3 Prediction based on Positive and Negative Probabilities(4 th model)	43
5.4.4 Additional Experiment	44
5.5 Performance Measure	45
CHAPTER VI RESULT AND DISCUSSION	47
6.1 Result	47
6.1.1 Seed	47
6.1.2 Supervised (1 st model)	47
6.1.2 Seeding with all prediction data (2 nd model)	48

6.1.3 Seeding based on deviant score (3 rd model)	48
6.1.4 Seeding based on positive and negative probabilities (4 th model)	49
6.2 Result from Additional Experiment	50
6.2.1 Result From 20% Initial Data	51
6.2.2 Result From 30% Initial Data	51
6.2.3 Result from 40% Initial Data	51
6.3 Discussion	51
6.3.1 Semi Supervised machine learning	52
6.3.2 Additional Experiment	54
CHAPTER VII CONCLUSIONS AND FUTURE RESEARCH	56
7.1 Conclusion	56
7.2 Future Research	57
REFERENCES	58

TABLE OF FIGURES

Figure 2.1 Result from comparing model by Triguero et al. (2013).....	8
Figure 3.1 Pseudo labelling (Kodžoman, 2017)	19
Figure 4.1 Multinomial Naïve Bayes	26
Figure 4.2 Multinomial Naïve Bayes implementing Seeding theory.....	27
Figure 5.1 Data frame of training data	37
Figure 5.2 Example of data	38
Figure 5.3 Implementation of Pre-processing.....	38
Figure 5.4 Result from Pre-processing	39
Figure 5.5 Implementation of Count Vectorizer	39
Figure 5.6 Word corpus from Count Vectorizer	39
Figure 5.7 Data after count Vectorizer.....	40
Figure 5.8 Implementation of Multinomial Naïve Bayes	40
Figure 5.9 Parameters used for machine learning.....	40
Figure 5.10 Set data for train and validation.....	41
Figure 5.11 Implementation of Seeding Theory	42
Figure 5.12 Implementation of seeding to add more data.....	42
Figure 5.13 Append pseudo label to future train data.....	43
Figure 5.14 Sampling from Deviant score	43
Figure 5.15 Sampling from probabilities	44
Figure 5.16 Implementation of Semi-supervised learning	45
Figure 5.17 Implementation of additional experiment.....	45
Figure 5.18 Implementation to evaluation	46
Figure 6.1 Result from sampling without filter.....	48
Figure 6.2 Result from Deviant score sampling	49
Figure 6.3 Result from probabilities sampling.....	50

TABLE OF TABLES

Table 2.1 Literature review	9
Table 3.1 Example of lowering letter.....	13
Table 3.2 Example of Stop word removal	13
Table 3.3 Example of Punctuation removal.....	13
Table 3.4 Example of stemming	14
Table 3.5 Confusion matrix	20
Table 4.1 Amount of test data	29
Table 4.2 Example of cleaning data.....	31
Table 4.3 Example of Count Vectorizer	31
Table 4.4 Probability result from Multinomial Naïve Bayes.....	32
Table 4.5 Parameter for Semi-Supervised	33
Table 6.1 Result from experiment.....	47
Table 6.2 Result from additional experiment.....	50
Table 6.3 Result from experiment.....	52
Table 6.4 Prediction table from seed model.....	53
Table 6.5 Comparing Supervised learning.....	54