

TABLE OF CONTENTS

APPROVAL PAGE	i
STATEMENT	ii
FOREWORDS	iii
TABLE OF CONTENTS	iv
LISTS OF TABLES	v
LISTS OF FIGURES	vi
CHAPTER I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	2
1.3 Research Objective	2
1.4 Research Scope	3
1.5 Research Benefits	3
CHAPTER II LITERATURE REVIEW	4
CHAPTER III BASIC THEORY	16
3.1 Weak Supervision	16
3.2 Weak Supervision with Rubrix Rulesets	16
3.2.1 Loading Data into Rubrix	17
3.2.2 Adding Rulesets with Rubrix	17
3.2.3 Labeling Model in Rubrix	18
3.2.4 Fitting with Snorkel in Rubrix	18
3.3 Text Preprocessing	19
3.3.1 Special Word Removal	19
3.3.2 Case Folding	19
3.3.3 Normalization	19
3.3.4 Stopwords Removal	20
3.3.5 Stemming	20
3.4 GloVe Embedding	20
3.5 Long short-term memory	21
3.6 CountVectorizer	22



3.7 Support Vector Machine	22
3.8 Performance Evaluation	22
3.8.1 Confusion Matrix	22
3.8.2 Accuracy	23
3.8.3 Recall	23
3.8.4 Precision	23
3.8.5 F1-Score	24
CHAPTER IV RESEARCH METHODOLOGY	25
4.1 Research Description	26
4.2 Tools and Materials	26
4.2.1 Tools	26
4.2.2 Materials	26
4.3 Data Collection	26
4.4 Weak Supervision	28
4.4.1 Weak Supervision with Rubrix Rulesets	28
4.4.2 Labeling Model in Rubrix	29
4.4.3 Fitting with Snorkel in Rubrix	29
4.5 GloVe Embedding	30
4.6 Classification	30
4.6.1 Support Vector Machine	30
4.6.2 Long short-term memory	31
4.7 Model Evaluation	31
CHAPTER V IMPLEMENTATION	32
5.1 Data Scraping Implementation	32
5.2 Preprocessing Implementation	32
5.2.1 Stopword Removal	32
5.2.2 Case Folding and Special Word Removal	33
5.2.3 Normalization	33
5.2.4 Stemming	34
5.3 Dataset Implementation	34
5.4 Weak Supervision	35



5.4.1 Loading data into Rubrix	35
5.4.2 Ruleset Labeling	36
5.5 Support Vector Machine	38
5.5.1 Mapping Labels and Training Data Prep Implementation	39
5.5.2 Pipeline Implementation	39
5.5.3 Testing Implementation	40
5.5.4 Prediction Implementation	40
5.6 LSTM with GloVe Embedding	40
5.6.1 GloVe Embedding	41
5.6.2 Data Splitting	41
5.6.3 Label Encoder and Tokenization	42
5.6.4 Pad Sequences	42
5.6.5 LSTM Modeling	43
5.6.6 Model Evaluation	43
CHAPTER VI RESULTS	44
6.1 Weekly Supervised Labeling Result	44
6.2 Data Result	44
6.3 Support Vector Machine Result	45
6.3.1 SVM Results with Manually Labeled Data	45
6.3.2 SVM Results with Weakly Supervised Data and Manually Labeled Testing Data	46
6.3.3 SVM Results with Combined Data	47
6.4 LSTM Result	47
6.4.1 LSTM Results with Manually Labeled Data	48
6.4.2 LSTM Results with Weakly Supervised Data and Manually Labeled Testing Data	51
6.4.3 LSTM Results with Combined Data	54
6.5 Comparison of Results	57
6.5.1 SVM Comparison	57
6.5.2 LSTM Comparison	58
CHAPTER VII CONCLUSION AND SUGGESTIONS	61



7.1 Conclusion

61

7.2 Suggestion

62

REFERENCES

63

APPENDICES

66

LIST OF TABLES

Table 2.1 Comparison between previous researches.....	9
Table 4.1 Sample of tweets from the dataset.....	27
Table 4.2 Total Tweets for each emotional label.....	27
Table 6.1 Weakly Supervised Labeling Score.....	44
Table 6.2 Data Values per label.....	45
Table 6.3 Parameters for SVM Model.....	45
Table 6.4 SVM Result with Manually Labeled Data.....	46
Table 6.5 SVM Result with Weakly Supervised Data and Manually Labeled Testing Data.....	46
Table 6.6 SVM Result with Combined Data.....	47
Table 6.7 Parameters for LSTM Model.....	48
Table 6.8 GloVe-LSTM with Manually Labeled Data.....	48
Table 6.9 GloVe-LSTM with Weakly Supervised Data and Manually Labeled Testing Data.....	51
Table 6.10 GloVe-LSTM with Combined Data.....	54
Table 6.11 End Result SVM Comparison.....	57
Table 6.12 Per-label Result SVM Comparison.....	58
Table 6.13 Per-label Result LSTM Comparison.....	59
Table 6.14 End Result LSTM Comparison.....	60

LIST OF FIGURES

Figure 3.1 Example of Rule Metrics in Rubrix.....	17
Figure 3.2 Diagram of Snorkel System.....	18
Figure 3.3 Example of Word Vectors.....	20
Figure 3.4 Diagram of Long short-term memory.....	21
Figure 3.5 Example of confusion matrix.....	23
Figure 4.1 Summary of Research flow diagram.....	25
Figure 4.2 The UI provided for Weak Labeling.....	29
Figure 4.3 Coverage of Rulesets in Rubrix.....	30
Figure 5.1 Twint Data Scraping Implementation.....	32
Figure 5.2 Stopword removal Implementation.....	33
Figure 5.3 Case Folding and Special word removal Implementation.....	33
Figure 5.4 Normalization Implementation.....	34
Figure 5.5 Stemming Implementation.....	34
Figure 5.6 Loading Unlabeled Data into Rubrix Implementation.....	34
Figure 5.7 Loading Labeled Data into Rubrix Implementation.....	35
Figure 5.8 Rubrix Reading pandas Dataset Implementation.....	35
Figure 5.9 Recording Unlabeled Data into Rubrix Implementation.....	36
Figure 5.10 Recording Labeled Data into Rubrix Implementation.....	36
Figure 5.11 Rubrix Ruleset Implementation.....	37
Figure 5.12 Rubrix Weak Labeling Implementation.....	37
Figure 5.13 Rubrix Weak Labeling Summary Implementation.....	38
Figure 5.14 Rubrix Performance of Weak Labeling Implementation.....	39
Figure 5.15 Mapping Labels and Training Data Prep Implementation.....	39
Figure 5.16 Pipeline Implementation.....	39
Figure 5.17 Testing Implementation.....	40
Figure 5.18 Prediction Implementation.....	40
Figure 5.19 GloVe Embedding Implementation.....	41
Figure 5.20 GloVe Embedding Matrix Implementation.....	41
Figure 5.21 Data Splitting Implementation.....	42



Figure 5.22 Label Encoder and Tokenization Implementation.....	42
Figure 5.23 Pad Sequences Implementation.....	42
Figure 5.24 Global Values Implementation.....	43
Figure 5.25 LSTM Modeling Implementation.....	43
Figure 5.26 Model Evaluation Implementation.....	43
Figure 6.1 Accuracy Graph of GloVe-LSTM Model with Manually Labeled Data.....	50
Figure 6.2 Loss Graph of GloVe-LSTM Model with Manually Labeled Data.....	51
Figure 6.3 Accuracy Graph of GloVe-LSTM Model with Weakly Supervised Data and Manually Labeled Data.....	53
Figure 6.4 Loss Graph of GloVe-LSTM Model with Weakly Supervised Data and Manually Labeled Data.....	53
Figure 6.5 Accuracy Graph of GloVe-LSTM Model with Combined Data.....	56
Figure 6.6 Loss Graph of GloVe-LSTM Model with Combined Data.....	56