

## TABLE OF CONTENTS

TITLE PAGE .....	i
VALIDATION PAGE .....	ii
STATEMENT .....	iii
PREFACE .....	iv
TABLE OF CONTENTS .....	v
TABLE LIST .....	vii
FIGURE LIST .....	viii
ABSTRACT .....	ix
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 Advantages .....	5
1.6 Proposal Organization .....	6
CHAPTER II LITERATURE REVIEW .....	7
2.1 Text Mining Growth Factors .....	7
2.2 Topic Modeling for Renewable Energy Act .....	9
2.3 Topic Modeling in Media Framing .....	10
2.4 Probabilistic Topic Modeling .....	12
2.5 Selection of the Optimal Number of Topics .....	13
CHAPTER III THEORY .....	17
3.1 Introduction to Topic Modeling .....	17
3.2 Topic modeling .....	17
3.3 Latent Dirichlet Allocation (LDA) .....	20
3.4 Negative Matrix Factorization (NMF) .....	25
3.5 Data Preprocessing .....	27
3.6 Data Filtering .....	28
3.7 Data Transformation .....	31
3.8 Topic Coherence Score .....	32
3.9 Document Embedding .....	33
3.10 t-distributed Stochastic Neighbor Embedding (t-SNE) .....	35
3.11 Orange .....	37
3.11.1 Corpus widget .....	43
3.11.2 Pre-process Text widget .....	44
3.11.3 Topic modeling widget .....	45
3.11.4 t-distributed Stochastic Neighbor Embedding Widget .....	46
CHAPTER IV RESEARCH METHODOLOGY .....	49
4.1 Thesis Workflow .....	49
4.2 Dataset Acquisition .....	50
4.3 Dataset from Kaggle .....	53
4.4 Orange Workflow .....	54
4.5 Tools and Materials .....	57
4.6 Workflow Diagram .....	57



CHAPTER V	IMPLEMENTATION .....	60
5.1	Algorithm Results .....	60
5.2	Topic Modeling Code .....	63
5.3	Custom Timer Code .....	65
5.4	Preprocessing Code .....	65
5.5	t-distributed Stochastic Neighbor Embedding Code .....	67
5.6	Visualization .....	68
5.7	Algorithm Advantages and Disadvantages .....	74
CHAPTER VI	CONCLUSION .....	75
6.1	Conclusion .....	75
REFERENCES	.....	78
APPENDIX	.....	80