

TABLE OF CONTENTS

TABLE OF CONTENTS	ii
ABSTRACT	v
I INTRODUCTION	1
1.1 Research Background	1
1.2 Research Problem	4
1.3 Research Objective	4
1.4 Research Scopes	5
1.5 Research Advantages	5
1.6 Research Schematic	6
II LITERATURE REVIEW	7
III FUNDAMENTAL THEORY	12
3.1 Natural Language Processing	12
3.2 Topic Modeling	13
3.3 Sentence Embedding using IndoSBERT	14
3.3.1 Sentence Embeddings and the Need for SBERT	14
3.3.2 SBERT Siamese Architecture	14
3.3.3 IndoBERT as Base Model	16
3.3.4 IndoSBERT Fine-tuning Process	16
3.4 Dimensionality Reduction using UMAP	17
3.5 Clustering using HDBSCAN	20
3.6 Topic Representation using c-TF-IDF	23
3.7 GSDMM for Pre-Cluster Knowledge Generation	24
3.8 BERTopic	26
3.9 G-BERTopic	27
3.10 Topic Coherence (C_v)	28
3.11 Davies-Bouldin Index	29
IV RESEARCH METHODOLOGY	31
4.1 Research Overview	31
4.2 Problem Analysis	33
4.3 Data Acquisition	34
4.4 Text Pre-processing	35
4.5 BERTopic Methodology	36
4.5.1 Sentence Embedding Generation	37
4.5.2 Dimensionality Reduction with UMAP	39
4.5.3 Clustering with HDBSCAN	40
4.5.4 Topic Representation with c-TF-IDF	42

4.6	G-BERTopic Methodology	43
4.6.1	Pre-Cluster Knowledge Generation	44
4.6.2	Embedding Generation and Feature Fusion	46
4.6.3	Clustering Refinement, Topic Representation, and Output	47
4.7	Evaluation Strategy	48
4.7.1	Evaluation Metrics	48
4.7.2	Model Comparison Strategy	49
4.8	Implementation	50
4.9	Report Writing	53
V	IMPLEMENTATION	54
5.1	Implementation Environment	54
5.2	Preparation	55
5.2.1	Research Setup	55
5.2.2	Dataset Preparation	56
5.3	Dataset Pre-Processing	59
5.3.1	Managing File Paths	59
5.3.2	Generating Raw Model Input	61
5.3.3	Cleaning Raw Model Input	66
5.3.4	Other Variation of Model Input	69
5.4	Sentence Embedding Implementation	70
5.5	Dimensionality Reduction Implementation	71
5.6	Clustering Methods	75
5.6.1	HDBSCAN Implementation	75
5.6.2	GSDMM Implementation	76
5.6.3	Agglomerative Clustering Implementation	78
5.7	Vectorizer Setup	79
5.8	cTF-IDF Implementation	80
5.9	BERTopic Implementation	80
5.10	G-BERTopic Implementation	83
5.11	Evaluation	87
5.11.1	Coherence Implementation	88
5.11.2	Davies–Bouldin Index Implementation	90
VI	RESULTS AND DISCUSSION	92
6.1	Evaluation Metric Result	92
6.2	Runtime Results	93
6.3	BERTopic and G-BERTopic Result	95
6.3.1	Topic Distribution in Embedding Space	95
6.3.2	Topic Word Distributions	97
6.3.3	Temporal Topic Evolution	101
6.4	Topic Modeling Results on Scoped News Data	104
6.5	Comparison of Results and Analysis	107

6.5.1	Evaluation Metrics	107
6.5.2	Computational Cost Comparison	109
6.5.3	Sensitivity Analysis on Maximum Topic Parameter in G-BERTopic	109
6.5.4	Topic Word Distributions	111
6.5.5	Temporal Frequency Behavior	112
6.6	Discussion	113
VI CONCLUSIONS AND SUGGESTIONS.		117
7.1	Conclusions	117
7.2	Suggestions	118
REFERENCES.		123