

## TABLE OF CONTENTS

HALAMAN PENGESAHAN.....	iii
PLAGIARISM STATEMENT .....	iv
ACKNOWLEDGEMENT .....	v
TABLE OF CONTENTS .....	vi
LIST OF TABLES .....	ix
LIST OF FIGURES .....	x
INTISARI.....	xii
ABSTRACT.....	xiii
CHAPTER I INTRODUCTION .....	1
1.1 Research Background.....	1
1.2 Research Problem.....	2
1.3 Research Scope .....	3
1.4 Research Objective.....	3
1.5 Research Benefit .....	3
1.6 Research Schematic.....	4
CHAPTER II LITERATURE REVIEW .....	5
CHAPTER III THEORETICAL BASIS.....	14
3.1 Data Ingestion .....	14
3.1.1 Stream Ingestion mode.....	14
3.1.2 Batch Ingestion Mode .....	15
3.2 Reddit as a Data Source .....	15
3.2.1 Pushshift Reddit Dataset .....	16
3.2.2 Official Reddit API .....	16
3.2.3 Pusher Real-time Reddit API.....	16
3.2.4 Pushshift Reddit API.....	17
3.3 Topic Modeling Technique .....	17
3.3.1 Latent Semantic Analysis (LSA) .....	17
3.3.2 Latent Dirichlet Allocation .....	18
3.4 Data Storage .....	19
3.5 Containerization .....	20
3.5.1 Docker .....	20
3.5.2 Airflow .....	22

3.5.3	Kafka .....	22
3.6	Evaluation Metrics .....	23
3.6.1	Latency .....	23
3.6.2	Coherence Score .....	23
3.6.3	Resource Utilization .....	24
CHAPTER IV	RESEARCH METHODOLOGY .....	25
4.1	Research Description .....	25
4.2	Model Initialization .....	26
4.2.1	Dataset .....	27
4.2.2	LDA .....	28
4.2.3	LSA .....	28
4.2.4	Saving Model and Push to MongoDB .....	29
4.3	Real-time Data Ingestion .....	29
4.3.1	Batch Pipeline .....	30
4.3.2	Stream Pipeline .....	31
4.4	Preprocess and Model Inference .....	31
4.4.1	LSA .....	31
4.4.2	LDA .....	32
4.5	Evaluation Metrics .....	32
4.5.1	Latency .....	32
4.5.2	Coherence Score .....	33
4.5.3	Resource Utilization .....	33
CHAPTER V	RESEARCH IMPLEMENTATION .....	35
5.1	Research Environment .....	35
5.2	Docker Setup .....	35
5.2.1	Airflow Setup for Batch Data Ingestion Pipeline .....	36
5.2.2	Kafka Setup for Streaming Pipeline .....	37
5.3	Importing libraries .....	39
5.4	Pre-processing .....	40
5.5	Initialization .....	41
5.5.1	Dataset .....	42
5.5.2	LDA .....	42
5.5.3	LSA .....	43
5.6	Batch Ingestion .....	45

5.6.1	LDA Model Inference .....	45
5.6.2	LSA Model Inference .....	48
5.7	Stream Ingestion .....	48
5.7.1	LDA Model Inference .....	49
5.7.2	LSA Model Inference .....	51
CHAPTER VI RESULT AND DISCUSSION .....		53
6.1	Initialization .....	53
6.1.1	MongoDB .....	53
6.1.2	Docker .....	53
6.1.3	Topic modeling .....	54
6.2	Batch Pipeline .....	54
6.2.1	LDA .....	55
6.2.2	LSA .....	57
6.3	Stream Pipeline .....	60
6.3.1	LDA .....	60
6.3.2	LSA .....	62
6.4	Performance Evaluation .....	65
6.4.1	Coherence Score .....	65
6.4.2	Latency .....	66
6.4.3	CPU Usage .....	67
6.4.4	Network Bandwidth .....	67
CHAPTER VII CONCLUSION .....		69
7.1	Conclusion .....	69
7.2	Future Work .....	69
REFERENCE .....		71

## LIST OF TABLES

<b>Table 2.1:</b> Literature Review .....	11
<b>Table 5.1:</b> Specification of the research computer .....	35
<b>Table 6.1:</b> Coherence Score and Latency for Batch Ingestion on the LDA Model .....	55
<b>Table 6.2:</b> CPU and Network Usage of Batch Ingestion on LDA Model .....	56
<b>Table 6.3:</b> Coherence Score and Latency for Batch Ingestion on LSA Model ....	58
<b>Table 6.4:</b> CPU and Network Usage of Batch Ingestion on LSA Model .....	58
<b>Table 6.5:</b> Coherence Score and Latency for Stream Ingestion on LDA Model .	60
<b>Table 6.6:</b> CPU and Network Usage of Stream Ingestion on LDA Model .....	61
<b>Table 6.7:</b> Coherence Score and Latency for Stream Ingestion on LSA Model ..	64
<b>Table 6.8:</b> CPU and Network Usage of Stream Ingestion on LSA Model .....	64
<b>Table 6.9:</b> Coherence Comparisons .....	66
<b>Table 6.10:</b> Latency Comparisons .....	66
<b>Table 6.11:</b> CPU Usage Comparisons .....	67
<b>Table 6.12:</b> Network Bandwidth Comparisons .....	67

## LIST OF FIGURES

<b>Figure 3.1:</b> Abstract typical view on a streaming analytics pipeline (Wingerath, Ritter & Gessert 2019) .....	14
<b>Figure 3.2:</b> LSA Topic Modeling (Albawi et al. 2020) .....	18
<b>Figure 3.3:</b> LDA Topic Modeling (Albawi et al. 2020) .....	18
<b>Figure 3.4:</b> Software delivery before and after Docker (Miell & Sayers 2019) .	21
<b>Figure 4.1:</b> Research Procedure Flow Chart .....	25
<b>Figure 4.2:</b> LDA Initialization .....	26
<b>Figure 4.3:</b> LSA Initialization .....	27
<b>Figure 4.4:</b> Batch Data Ingestion Pipeline .....	29
<b>Figure 4.5:</b> Streaming Data Ingestion Pipeline .....	30
<b>Figure 5.1:</b> Batch Pipeline Setup from <i>Dockerfile</i> .....	36
<b>Figure 5.2:</b> Batch Pipeline Setup from <i>docker-compose.yml</i> .....	37
<b>Figure 5.3:</b> Streaming Pipeline Docker Setup from <i>requirements.txt</i> .....	37
<b>Figure 5.4:</b> Streaming Pipeline Docker Setup from <i>docker-compose.yml</i> .....	38
<b>Figure 5.5:</b> Streaming Pipeline Docker Setup from <i>Dockerfile</i> .....	38
<b>Figure 5.6:</b> Streaming Pipeline Docker Setup from <i>start.sh</i> .....	39
<b>Figure 5.7:</b> Libraries Necessary for LDA Model Processing .....	39
<b>Figure 5.8:</b> Libraries Necessary for LSA Topic Model Processing .....	40
<b>Figure 5.9:</b> Pre-processing Function .....	40
<b>Figure 5.10:</b> Reddit Scraping for Initial Training Dataset .....	41
<b>Figure 5.11:</b> LDA Topic Modeling Initialization .....	43
<b>Figure 5.12:</b> LSA Topic Modeling Initialization Implementation .....	45
<b>Figure 5.13:</b> Batch Ingestion for LDA Implementation .....	46
<b>Figure 5.14:</b> Batch Ingestion for LDA Implementation .....	47
<b>Figure 5.15:</b> Streaming and Comparing new Data in Kafka Produce .....	49
<b>Figure 5.16:</b> Kafka Consumer to Infer LDA Model .....	50
<b>Figure 5.17:</b> Kafka Consumer to Infer LSA Model .....	52
<b>Figure 6.1:</b> LDA Topic Model Coherence Score Records for Batch Ingestion ...	54

<b>Figure 6.2:</b> Batch Data Ingestion Latency in LDA Topic Modeling .....	55
<b>Figure 6.3:</b> CPU Usage for Batch Data Ingestion in LDA Topic Modeling.....	56
<b>Figure 6.4:</b> Network Consumption in Batch Ingestion for LDA Topic Modeling .....	57
<b>Figure 6.5:</b> LSA Topic Model Coherence Score Records for Batch Ingestion....	57
<b>Figure 6.6:</b> Batch Data Ingestion Latency in LSA Topic Modeling .....	58
<b>Figure 6.7:</b> Network Consumption in Batch Ingestion for LSA Topic Modeling	59
<b>Figure 6.8:</b> CPU Usage for Batch Data Ingestion in LSA Topic Modeling .....	59
<b>Figure 6.9:</b> LDA Topic Model Coherence Score Records for Stream Ingestion .	60
<b>Figure 6.10:</b> Stream Data Ingestion Latency in LDA Topic Modeling .....	61
<b>Figure 6.11:</b> Network Bandwidth for Stream Data Ingestion in LDA Topic Modeling .....	62
<b>Figure 6.12:</b> CPU Usage for Stream Data Ingestion in LDA Topic Modeling....	62
<b>Figure 6.13:</b> Stream Data Ingestion Coherence Score in LSA Topic Modeling..	63
<b>Figure 6.14:</b> Stream Data Ingestion Latency in LSA Topic Modeling .....	63
<b>Figure 6.15:</b> CPU Usage for Stream Data Ingestion in LSA Topic Modeling ....	64
<b>Figure 6.16:</b> Network Consumption in Stream Ingestion for LSA Topic Modeling .....	65