Topic Modelling on Dark Web Forums using LDA

Submitted Towards the Partial Fulfilment of the Requirements
for the Award of the Degree of

Bachelor of Technology

by

Vidur Dua	Roll No. 23124119
Shivansh Dutta	Roll No. 23124102
Vaibhav Kanda	Roll No. 23124113
Karman Singh	Roll No. 22124054

Under the Mentorship of

Dr. Jaspal Kaur Saini
Assistant Professor - IT dept.

Department of Information Technology
Dr B R Ambedkar National Institute of Technology
Jalandhar-144008, Punjab (INDIA)

Undertaking

We hereby declare that the project work presented in this report, entitled **"Topic Modeling of Dark Web Forums Using LDA"**, submitted to the Department of Information Technology, Dr. B R Ambedkar National Institute of Technology, Jalandhar, is our original work.

This work is carried out as part of the **Minor Project** requirements of the Bachelor of Technology programme in Information Technology. We further declare that this work has not been plagiarized or submitted, either in part or full, for the award of any other degree or diploma at this or any other institute.

In the event that any part of this declaration is found to be incorrect, we fully understand that appropriate academic action may be taken.

Vidur Dua (Roll No. 23124119)

Shivansh Dutta (Roll No. 23124102)
Vaibhav Kanda (Roll No. 23124113)

Karman Singh (Roll No. 22124054)

Department of Information Technology
Dr B R Ambedkar National Institute of Technology
Jalandhar, Punjab, India
November 2025

Certificate

This is to certify that the project report entitled **"Topic Modeling of Dark Web Forums Using LDA"**, submitted by Vidur Dua (Roll No. 23124119), Shivansh Dutta (Roll No. 23124102), Vaibhav Kanda (Roll No. 23124113), Karman Singh (Roll No. 22124054) to Dr B R Ambedkar National Institute of Technology, Jalandhar, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Information Technology, has been carried out under our supervision.

We also certify that this work is original and has not been submitted elsewhere for the award of any degree or diploma.

Dr. Jaspal Kaur Saini

Assistant Professor

Department of Information Technology
Dr B R Ambedkar National Institute of Technology
Jalandhar, Punjab, India
November, 2025

Acknowledgement

We would like to express our sincere gratitude to all those who have contributed to the successful completion of this project.

First and foremost, we extend our heartfelt thanks to our project supervisor, Dr. Jaspal Kaur Saini, for her invaluable guidance, continuous support, and encouragement throughout the duration of this project. Her expertise and constructive feedback have been instrumental in shaping this work.

We are deeply grateful to Dr. Vijay Kumar, Head of the Department of Information Technology, for providing us with the necessary facilities and resources to carry out this project.

We would also like to thank all the faculty members of the Department of Information Technology for their support and valuable suggestions during various stages of this project.

Our sincere thanks go to our families and friends for their constant encouragement and support throughout our academic journey.

Finally, we acknowledge the contributions of all those who directly or indirectly helped us in completing this project successfully.

Vidur Dua (Roll No. 23124119)
Shivansh Dutta (Roll No. 23124102)
Vaibhav Kanda (Roll No. 23124113)
Karman Singh (Roll No. 22124054)

Abstract

This project presents a comprehensive analysis of seven dark web and underground Islamic discussion forums using Latent Dirichlet Allocation (LDA)–based topic modeling. With the increasing volume of unstructured text generated on hidden online platforms, there is a growing need for automated techniques to identify underlying themes, behavioral patterns, and ideological trends. Manual inspection of such large datasets is both impractical and inefficient, making topic modeling an essential tool for scalable analysis.

The primary objective of this work is to preprocess seven forum datasets, construct baseline LDA models for each, and interpret the thematic structures emerging from these platforms. The methodological pipeline includes extensive data cleaning, tokenization, lemmatization, stop-word removal, dictionary and corpus creation, and the training of 10-topic LDA models. Visualization tools such as pyLDAvis, word clouds, and topic distribution charts are employed to explore inter-topic distances, keyword significance, and dominant discussion patterns.

Experimental results show distinct thematic variations across the forums. Ansar1 exhibits conflict-centric and militant narratives; Gawaher emphasizes spiritual guidance and community bonding; TurnToIslam displays multilingual, lifestyle-oriented, and religious discussions; IslamicNetwork and MyIWC present balanced mixes of religious, social, and political discourse; while the remaining datasets demonstrate unique combinations of devotional, geopolitical, and conversational themes. These variations highlight the linguistic diversity, ideological tendencies, and sociocultural behavior present across different online communities.

Overall, this study demonstrates that LDA is an effective and scalable method for uncovering meaningful patterns within noisy, real-world forum datasets. The insights gained provide a foundation for future research in threat intelligence, extremist behavior monitoring, digital sociology, and advanced semantic topic modeling.

Keywords: Topic Modeling, LDA, Dark Web Forums, NLP, Text Mining, Semantic Analysis

Contents

Acknowledgement **4**

Abstract **5**

List of Figures **9**

List of Tables **10**

1 Introduction **11**
1.1 Overview . . . 11
1.2 Problem Statement . . . 11
1.3 Motivation . . . 12
1.4 Objectives . . . 12
1.5 Scope of the Project . . . 12
1.6 Organization of Report . . . 13

2 Literature Review **14**
2.1 Introduction . . . 14
2.2 Existing Systems/Approaches . . . 14
2.2.1 Traditional Approaches . . . 14
2.2.2 Modern Solutions . . . 14
2.3 Technology Review . . . 15
2.3.1 Latent Dirichlet Allocation (LDA) . . . 15
2.3.2 Machine Learning and Deep Learning Methods . . . 15
2.4 Comparative Analysis . . . 16
2.5 Research Gaps . . . 16
2.6 Summary . . . 17

3 Background and Preliminaries **18**
3.1 Introduction . . . 18
3.2 Fundamental Concepts . . . 18
3.2.1 Natural Language Processing (NLP) . . . 18
3.2.2 Topic Modeling . . . 18

3.3 Technologies and Tools . . . 19
3.3.1 Technology Stack . . . 19
3.3.2 Development Environment . . . 20
3.4 Theoretical Framework . . . 20
3.4.1 Bag-of-Words Model . . . 20
3.4.2 Latent Dirichlet Allocation (LDA) . . . 20
3.4.3 Coherence Score . . . 21
3.5 Summary . . . 21

4 Proposed Methodology **22**
4.1 Introduction . . . 22
4.2 System Overview . . . 22
4.3 System Architecture . . . 22
4.3.1 Architecture Design . . . 22
4.3.2 Component Description . . . 23
4.4 Module Description . . . 23
4.4.1 Module 1: Data Preprocessing . . . 23
4.4.2 Module 2: Corpus and Dictionary Creation . . . 24
4.4.3 Module 3: LDA Topic Modeling . . . 24
4.4.4 Module 4: Topic Visualization . . . 24
4.5 Algorithms and Techniques . . . 25
4.5.1 Latent Dirichlet Allocation (LDA) . . . 25
4.6 Design Diagrams . . . 25
4.6.1 Class Diagram . . . 25
4.6.2 Activity Diagram . . . 26
4.6.3 Use Case Diagram . . . 26
4.6.4 Sequence Diagram . . . 27
4.6.5 State Diagram . . . 28
4.6.6 Architectural Design Patterns . . . 29
4.6.7 Summary . . . 30
4.7 Database Design . . . 31
4.8 User Interface Design . . . 31
4.9 Security and Privacy . . . 31
4.10 Implementation Strategy . . . 31
4.11 Summary . . . 31

5 Experimental Results and Analysis **32**
5.1 Introduction . . . 32
5.2 Experimental Setup . . . 32

5.2.1 Hardware Configuration 32
5.2.2 Software Configuration 33
5.3 Implementation Details 33
5.4 Experimental Results 33
5.4.1 Experimental Results – Ansar1 Dataset 33
5.4.2 Experimental Results – MyIWC Dataset 36
5.4.3 Experimental Results – Gawaher Dataset 37
5.4.4 Experimental Results – TurnToIslam Dataset 39
5.4.5 Experimental Results – IslamicNetwork Dataset 41
5.4.6 Experimental Results – Islamic Awakening Dataset 43
5.4.7 Experimental Results – Ummah Dataset 46
5.5 Visualization-Based Insights 49
5.6 Analysis and Discussion 49
5.6.1 Key Findings 49
5.6.2 Advantages 50
5.6.3 Limitations 50
5.7 Summary 50

6 Conclusion and Future Work **51**
6.1 Conclusion 51
6.1.1 Summary of Work 51
6.1.2 Key Contributions 52
6.1.3 Achievement of Objectives 52
6.1.4 Impact and Significance 52
6.2 Future Scope 52
6.2.1 Short-term Enhancements 53
6.2.2 Long-term Vision 53
6.2.3 LDA Improvement and Advanced Model Search 53
6.2.4 Research Directions 54
6.3 Final Remarks 54

References **55**

List of Figures

4.1 Class Diagram for the Topic Modeling Pipeline 25
4.2 Activity Diagram of the Topic Modeling Workflow 26
4.3 Use Case Diagram Showing Researcher and User Interactions 27
4.4 Sequence Diagram for Topic Modeling Execution (Researcher POV) . . . 27
4.5 Sequence Diagram for Topic Modeling Execution (User POV) 28
4.6 State Diagram Representing System Execution States 29
4.7 Three-Layered Architecture Pattern 30
4.8 Strategy Pattern 30

5.1 Topic Distribution Across Ansar1 Forum 35
5.2 Ansar1 pyLDAvis 35
5.3 Topic Distribution Across MyIWC Forum 37
5.4 MyIWC pyLDAvis 37
5.5 Topic Distribution Across Gawaher Forum 39
5.6 Gawaher pyLDAvis 39
5.7 Topic Distribution Across TurnToIslam Forum 41
5.8 TurnToIslam pyLDAvis 41
5.9 Topic Distribution Across IslamicNetwork Forum 43
5.10 IslamicNetwork pyLDAvis 43
5.11 Topic Distribution Across IslamicAwakening Forum 45
5.12 IslamicAwakening pyLDAvis 46
5.13 Topic Distribution Across Ummah Forum 48
5.14 Ummah pyLDAvis 49

List of Tables

2.1 Comparison of Existing Approaches 16

5.1 Ansar1 Topic–Label Mapping 34
5.2 MyIWC Topic–Label Mapping 36
5.3 Gawaher Topic–Label Mapping 38
5.4 TurnToIslam Topic–Label Mapping 40
5.5 IslamicNetwork Topic–Label Mapping 42
5.6 Islamic Awakening Topic–Label Mapping 44
5.7 Ummah Topic–Label Mapping (Simulated) 47

Chapter 1

Introduction

1.1 Overview

This project investigates thematic patterns within dark web and underground Islamic forums using Latent Dirichlet Allocation (LDA). These forums contain large volumes of unstructured, noisy, and linguistically diverse text, making manual inspection extremely challenging. Topic modeling provides an effective way to automatically extract meaningful themes, uncover hidden discourse structures, and understand the narrative focus of different online communities.

In this work, seven datasets—TurnToIslam, Ansar1, MyIWC, IslamicNetwork, Gawaher, and two additional Islamic discussion forums—are analyzed to examine linguistic diversity, ideological content, social conversations, and community-driven discussions. By applying a standardized natural language processing (NLP) workflow across all datasets, this study highlights how religious, political, and conversational themes differ across platforms.

1.2 Problem Statement

Dark web and underground forums host diverse religious, political, and ideological discussions. Because the data is unstructured, multilingual, and often noisy, extracting meaningful insights manually is not feasible. Existing studies typically focus on a single forum or limited datasets, leading to insufficient comparative understanding of community behaviors.

The key issues addressed in this project are:

- The lack of automated approaches for analyzing large volumes of dark-web forum data.
- Difficulty in identifying latent themes due to noise, informal language, and multilingual content.
- Limited comparative studies across multiple Islamic and dark-web forums.
- Need for scalable topic modeling techniques to capture thematic variations across platforms.

This project applies LDA to overcome these challenges and generate interpretable thematic structures across seven forums.

1.3 Motivation

Understanding the content of underground and religious forums is essential for research in cybersecurity, social behavior, extremism analysis, and digital communication. These platforms influence user narratives, ideological development, and social interactions. Analyzing their discussions reveals how information spreads, how communities respond to global events, and how certain themes dominate specific platforms.

Key motivations include:

- To systematically analyze large-scale forum data that cannot be interpreted manually.
- To compare thematic variations across seven Islamic and dark-web forums.
- To apply LDA as a scalable topic modeling approach for unstructured text.
- To contribute insights useful for threat intelligence, religious discourse analysis, and sociopolitical research.

1.4 Objectives

The primary objectives of this project are:

1. To preprocess and clean seven forum datasets using a uniform NLP pipeline.
2. To train 10-topic LDA models independently for each dataset.
3. To interpret and label topics using keyword distributions and representative documents.
4. To compare thematic structures across forums and identify unique and shared discourse patterns.

1.5 Scope of the Project

The scope of this project includes:

- Data preprocessing using tokenization, stop-word removal, and lemmatization.
- Dictionary and corpus creation using gensim.
- Training of LDA models with 10 topics per dataset.
- Visualization and analysis using pyLDAvis, word clouds, and topic distribution plots.

The project does not include sentiment analysis, temporal trend analysis, neural topic modeling (e.g., BERTopic), or predictive analytics.

1.6 Organization of Report

This report is organized as follows:

- **Chapter 2** provides background theory for LDA and related NLP concepts.
- **Chapter 3** presents an in-depth literature review of prior research.
- **Chapter 4** describes the methodology, preprocessing workflow, and LDA model configuration.
- **Chapter 5** provides experimental results and dataset-wise topic analysis.
- **Chapter 6** discusses the comparative findings, conclusions, and future research directions.

Chapter 2

Literature Review

2.1 Introduction

The purpose of this literature review is to examine existing research, methodologies, and computational approaches used to analyze dark web forums, extremist communication networks, and large-scale unstructured text datasets. This review highlights traditional manual methods as well as modern machine learning and topic modeling techniques that support the extraction of latent patterns, behavioral cues, and thematic structures from online discussion platforms. The selected works provide foundational context for understanding how automated methods—particularly Latent Dirichlet Allocation (LDA)—can be applied to multi-forum datasets in the context of extremist and Islamic online communities.

2.2 Existing Systems/Approaches

Research on dark web forums has evolved over two decades, beginning with manual information collection and progressing toward automated content analysis, supervised learning, and deep learning-based detection systems.

2.2.1 Traditional Approaches

Early studies focused on manual inspection and case-based analysis of extremist forums. Chen et al. conducted one of the earliest and most influential investigations into Jihadist websites, introducing a semi-automated methodology for collecting, filtering, and analyzing dark web information [1]. Their work highlighted the challenges of multilingual data, information overload, and fragmented content across extremist platforms.

Similarly, Abbasi et al. explored multilingual sentiment classification on extremist websites, demonstrating the difficulty of analyzing heterogeneous forum data using rule-based and linguistic techniques [3]. These traditional studies relied heavily on human interpretation, making them unsuitable for large-scale datasets.

2.2.2 Modern Solutions

Modern research introduced automated text mining, supervised learning, and topic modeling to address scalability.

Scanlon and Gerber proposed a supervised machine learning approach to detect cyber-recruitment activities in extremist forums using Naïve Bayes, logistic regression, and SVM classifiers [4]. Their work demonstrated the feasibility of identifying recruitment-oriented posts automatically.

Saini and Bansal expanded this domain by analyzing weapon procurement discussions on four dark web forums using machine learning classification techniques [5]. Their study highlighted the increasing trend of extremist groups using online platforms for illegal purchases.

Deep learning advancements further enhanced detection capabilities. Saini introduced an LSTM-based framework for identifying violent activities such as recruitment and weapon procurement, achieving higher accuracy compared to classical machine learning models [6].

Recent advancements in topic modeling, including Semantic-LDA and improved coherence-driven variants, have been proposed to increase accuracy, reduce topic overlap, and enhance semantic richness [7]. These improvements directly support scalable multi-forum topic extraction.

2.3 Technology Review

This section reviews the key computational methods and algorithms relevant to the proposed system, specifically topic modeling, NLP preprocessing, and LDA-based techniques.

2.3.1 Latent Dirichlet Allocation (LDA)

LDA is an unsupervised probabilistic topic modeling method introduced to discover latent patterns in large textual corpora. It models documents as mixtures of topics and topics as distributions of words, making it suitable for unstructured data. LDA has been used extensively in dark web forum studies, including Chen et al.'s Jihad Web analysis [1] and multiple works on extremist communities.

Core steps include:

- Dictionary and corpus construction
- Gibbs sampling or variational inference
- Topic assignment and interpretation

2.3.2 Machine Learning and Deep Learning Methods

Traditional ML classifiers such as SVM, logistic regression, and Naïve Bayes have been used for recruitment and weapon procurement detection [4,5]. Deep learning approaches,

particularly LSTM architectures, outperform classical models in sequence-based text classification tasks due to their ability to capture long-term dependencies [6]. Although these supervised approaches align with extremist content detection, they require labeled datasets, which limits scalability.

2.4 Comparative Analysis

A comparison of major approaches from the literature is provided in Table 2.1.

Table 2.1: Comparison of Existing Approaches

Approach	**Advantages**	**Limitations**	**Reference**
Manual Dark Web Analysis	High interpretability	Not scalable; time-consuming	[1]
ML-based Recruitment Detection	Good classification accuracy	Requires labeled data	[4]
Weapon Procurement Detection	Domain-specific insights	Limited to specific forums	[5]
LSTM-based Violent Activity Detection	High accuracy; deep features	Computationally expensive	[6]
Improved LDA / Semantic-LDA	Better coherence; low overlap	Higher computational cost	[7]

2.5 Research Gaps

Despite extensive work in extremist forum analysis, several gaps remain:

- Limited studies conduct **multi-forum comparative topic modeling** across more than 3–4 datasets.
- Most existing research relies on **supervised learning**, which requires annotated data.
- Few studies explore **unsupervised topic modeling** as the primary analytical method for Islamic and dark web forums.
- Topic overlap, semantic redundancy, and low coherence remain major limitations in traditional LDA.
- Lack of standardized preprocessing across multilingual forums introduces inconsistency.

2.6 Summary

The reviewed literature demonstrates a clear evolution from manual dark web analysis toward automated topic modeling, machine learning, and deep learning techniques. Existing work highlights the importance of understanding extremist communication patterns but is often constrained by limited datasets, reliance on labeled data, or domain-specific scope.

This project addresses these gaps by applying a unified NLP pipeline and LDA topic modeling across seven forum datasets, enabling comparative thematic analysis at scale. The insights gained contribute to text mining research, dark-web behavior analysis, and future development of improved topic modeling techniques.

Chapter 3

Background and Preliminaries

3.1 Introduction

This chapter provides the essential background knowledge required to understand the methods and analytical approaches used in this project. Since the objective of this work is to perform topic modeling across seven different dark web and Islamic forum datasets, it is crucial to understand the core concepts of Natural Language Processing (NLP), topic modeling, Latent Dirichlet Allocation (LDA), and the preprocessing workflows that make large-scale text mining possible. This chapter also describes the technologies, tools, and theoretical foundations that support dataset processing and model development.

3.2 Fundamental Concepts

3.2.1 Natural Language Processing (NLP)

Natural Language Processing is a subfield of artificial intelligence concerned with enabling computers to understand, interpret, and generate human language. NLP techniques form the backbone of this project by enabling text cleaning, tokenization, lemmatization, stop-word removal, and corpus construction.

Key NLP preprocessing operations include:

- **Tokenization** — breaking text into individual words or tokens.
- **Lemmatization** — reducing words to their base or dictionary form.
- **Stop-word removal** — eliminating frequently occurring but semantically weak words such as "the," "and," "is."
- **Dictionary creation** — mapping unique words to integer IDs.
- **Corpus building** — forming a bag-of-words representation of documents.

These steps ensure that the dataset is clean, consistent, and ready for topic modeling.

3.2.2 Topic Modeling

Topic modeling refers to a family of statistical methods used to discover abstract topics from large text collections. It is an unsupervised approach, meaning it does not require

labeled data. Topic modeling is particularly suitable for analyzing dark web forums because the data is unstructured, noisy, and multilingual.

The goal of topic modeling is to:

- Identify dominant themes in the dataset.
- Represent documents as mixtures of topics.
- Understand patterns of user discussions across forums.

LDA, one of the most widely used topic modeling algorithms, is central to the methodology used in this project.

3.3 Technologies and Tools

3.3.1 Technology Stack

For this project, the following tools and technologies were used:

- **Programming Language:** Python Widely used for NLP and machine learning due to libraries such as gensim, spaCy, and nltk.
- **Frameworks and Libraries:**
  - **gensim** — for implementing LDA, dictionary creation, and corpus modelling.
  - **spaCy** — for tokenization, lemmatization, and advanced NLP preprocessing.
  - **nltk** — for stop-word lists and preprocessing utilities.
  - **pyLDAvis** — for interactive visualization of topics and inter-topic distances.
  - **matplotlib / wordcloud** — for visual representation of topic results.
- **Dataset Handling:** Python's `pandas` and `csv` modules were used for cleaning, formatting, and structuring the seven text datasets.
- **Hardware/Runtime Tools:** Jupyter Notebook and Google Colab were used for experimentation and debugging due to their interactive environment.

3.3.2 Development Environment

The development environment consisted of:

- **IDE:** Jupyter Notebook / Google Colab for stepwise execution and visualization.
- **Version Control:** Git/GitHub for source code management and maintaining reproducibility.
- **Python Environment:** Python 3.10+ with virtual environments for dependency control.

This environment ensured a modular workflow and reproducible experiments across all datasets.

3.4 Theoretical Framework

3.4.1 Bag-of-Words Model

The bag-of-words (BoW) model represents text as the frequency distribution of words. BoW ignores grammar and word order but provides a simple, effective representation for topic modeling.

3.4.2 Latent Dirichlet Allocation (LDA)

LDA is a generative probabilistic model where:

- Each document is a mixture of latent topics.
- Each topic is a distribution of words.

The mathematical intuition behind LDA:

- A Dirichlet prior is applied to document–topic distributions.
- Another Dirichlet prior is applied to topic–word distributions.
- Gibbs sampling or variational inference is used to approximate posterior probabilities.

LDA is particularly suitable for:

- Large-scale text corpora.
- Unsupervised pattern discovery.
- Multi-forum thematic comparison.

3.4.3 Coherence Score

Topic coherence measures the semantic consistency of the generated topics. Higher coherence indicates more interpretable topics. This metric is essential for evaluating LDA models and selecting the number of topics.

3.5 Summary

This chapter reviewed the foundational concepts needed to understand the topic modeling process used in this project. It introduced NLP preprocessing techniques, LDA modeling principles, and the technologies/tools required for dataset handling and topic visualization. These concepts form the backbone of the methodology applied in the subsequent chapters.

Chapter 4

Proposed Methodology

4.1 Introduction

This chapter presents the proposed methodology for performing topic modeling and comparative analysis on seven dark web and Islamic forum datasets. The methodology integrates Natural Language Processing (NLP), text preprocessing, Latent Dirichlet Allocation (LDA), and visualization techniques. The goal is to extract meaningful topics from each dataset, compare thematic structures across forums, and understand the linguistic and ideological patterns present within these communities.

4.2 System Overview

The proposed system consists of a multi-stage pipeline that begins with dataset loading and cleaning, followed by text preprocessing, dictionary and corpus creation, LDA model training, and finally topic interpretation and visualization. Each stage contributes to systematically transforming noisy raw forum text into interpretable and structured thematic outputs.

4.3 System Architecture

The architecture is designed as a sequential processing workflow using NLP and topic modeling techniques. Each stage is modular, allowing flexibility in processing different datasets while maintaining a uniform methodology.

4.3.1 Architecture Design

The architecture consists of the following layers:

- **Dataset Input Layer** — Loads raw text data from seven forum datasets.
- **Preprocessing Layer** — Converts raw text into a clean, structured format.
- **Corpus Construction Layer** — Creates dictionaries and bag-of-words representations.
- **Topic Modeling Layer** — Applies LDA to extract latent topics.

- **Visualization Layer** — Generates topic distributions, word clouds, and pyLDAvis plots.
- **Comparative Analysis Layer** — Compares topics across datasets.

4.3.2 Component Description

Dataset Processing Component

This component loads and standardizes input datasets. It handles variations in formatting, special characters, HTML tags, and multilingual content.

Preprocessing Component

This module performs tokenization, lemmatization, stop-word removal, and normalization to prepare text for modeling.

LDA Modeling Component

This component trains a 10-topic LDA model for each dataset using gensim's implementation.

Visualization Component

Generates outputs such as:

- pyLDAvis inter-topic distance maps
- Word clouds for each topic
- Topic weights and keyword importance

4.4 Module Description

4.4.1 Module 1: Data Preprocessing

Purpose: Convert raw text into a clean, machine-readable format.

Functionality:

- Tokenization of forum posts
- Lemmatization using spaCy NLP pipeline
- Removal of stop-words, punctuation, and noise
- Handling multi-language content

Implementation Details: Implemented in Python using spaCy and nltk. Special text patterns such as URLs, repeated characters, and HTML tags are removed using regex operations.

4.4.2 Module 2: Corpus and Dictionary Creation

Purpose: Transform preprocessed tokens into numerical structures for LDA.

Functionality:

- Creating a gensim dictionary mapping words to integer IDs
- Constructing a Bag-of-Words (BoW) corpus
- Filtering extreme tokens (too common or too rare)

Implementation Details: Uses gensim's built-in dictionary and corpus tools.

4.4.3 Module 3: LDA Topic Modeling

Purpose: Discover hidden thematic structures in the datasets.

Functionality:

- Training 10-topic LDA models for each dataset
- Extracting topic-word distributions
- Identifying dominant topics per dataset

Implementation Details: Uses gensim's LdaModel with optimized parameters:

- Number of topics = 10
- Passes = 20
- Alpha = 'auto'
- Beta (eta) = 'auto'

4.4.4 Module 4: Topic Visualization

Purpose: Present topic outputs in interpretable graphical form.

Functionality:

- pyLDAvis topic maps
- Word clouds for each topic
- Distribution charts for dominant topics

4.5 Algorithms and Techniques

4.5.1 Latent Dirichlet Allocation (LDA)

LDA is the primary algorithm used in this project. It assumes:

- Documents are mixtures of latent topics.
- Topics are distributions of words.

The generative process:

1. Choose a topic distribution for each document.
2. For each word in the document:
    - Select a topic.
    - Draw a word from the topic's distribution.

4.6 Design Diagrams

Design diagrams are used to visually represent the structure, behavior, and workflow of the proposed topic modeling system. These diagrams help in understanding how different components interact, how data flows through the system, and how users and researchers engage with the analysis pipeline. All diagrams follow standard UML and software architecture conventions.

4.6.1 Class Diagram

The Class Diagram, presented in Figure 4.1, represents the static structure of the system. It illustrates the key classes responsible for preprocessing, dictionary creation, corpus generation, LDA model training, and visualization. Relationships between classes highlight the modular and extensible design of the system.

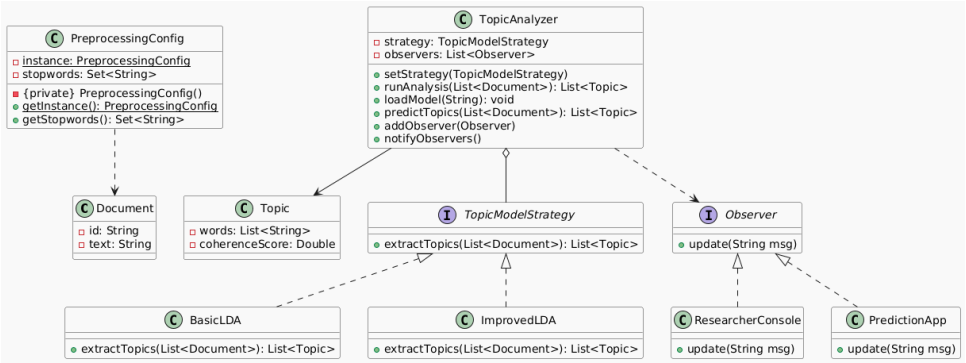

Figure 4.1: Class Diagram for the Topic Modeling Pipeline

4.6.2 Activity Diagram

Figure 4.2 shows the Activity Diagram, which depicts the dynamic workflow of the system. It models the sequence of activities from dataset loading and preprocessing to topic modeling and result visualization, including decision points and execution flow.

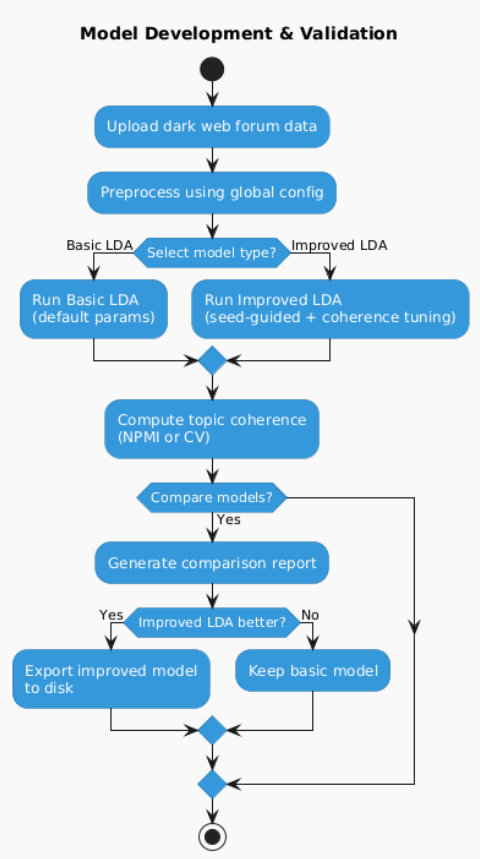

Figure 4.2: Activity Diagram of the Topic Modeling Workflow

4.6.3 Use Case Diagram

The Use Case Diagram shown in Figure 4.3 represents the interaction between the system and its actors. Two actors are identified: the *Researcher* and the *User*. The diagram captures use cases such as dataset selection, model execution, visualization generation, and result interpretation.

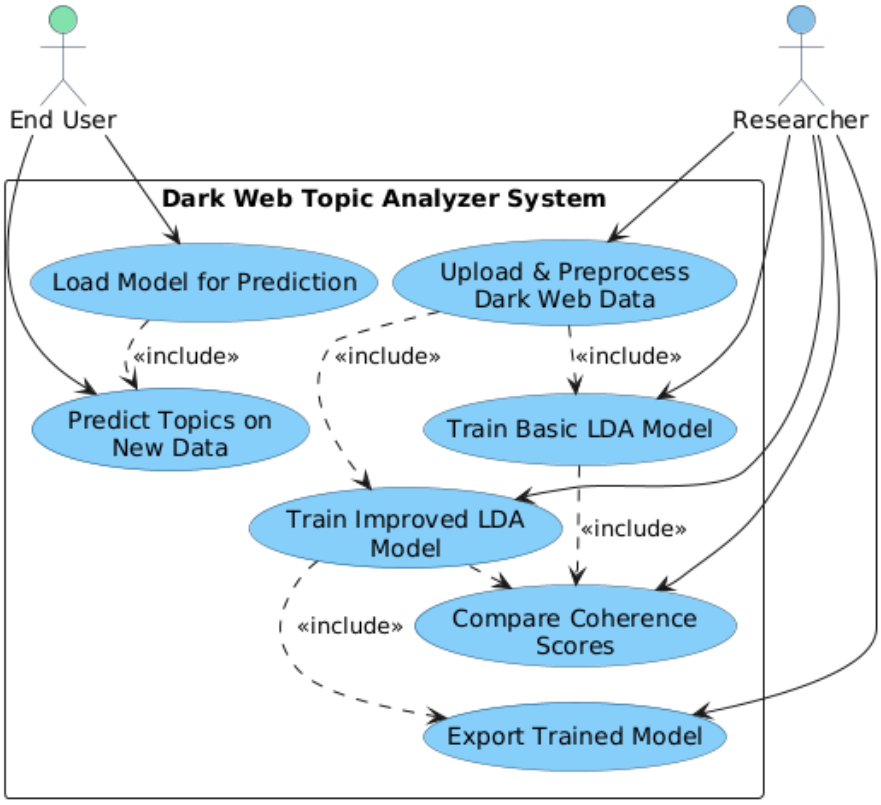

Figure 4.3: Use Case Diagram Showing Researcher and User Interactions

4.6.4 Sequence Diagram

The Sequence Diagram, illustrated in Figure 4.5, shows the temporal order of interactions between system components. Separate flows are modeled from both researcher and user perspectives, demonstrating how control passes from data ingestion to preprocessing, modeling, visualization, and output generation.

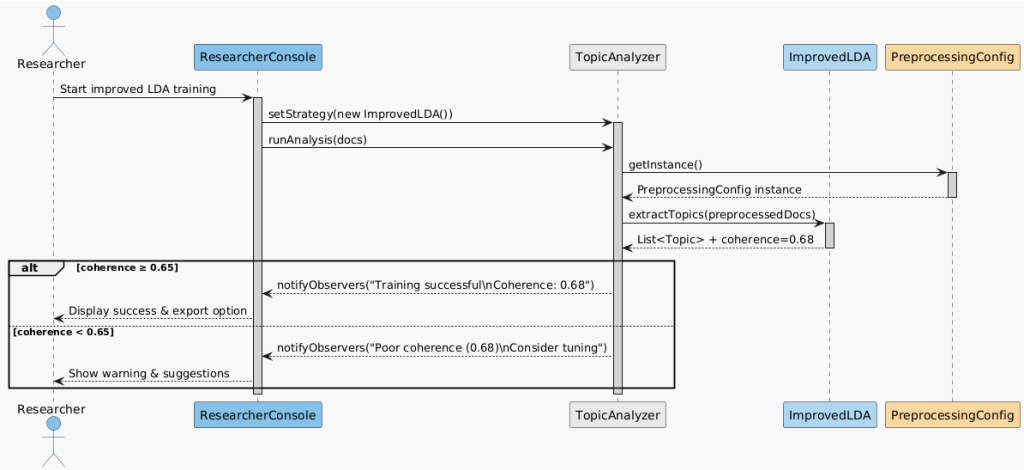

Figure 4.4: Sequence Diagram for Topic Modeling Execution (Researcher POV)

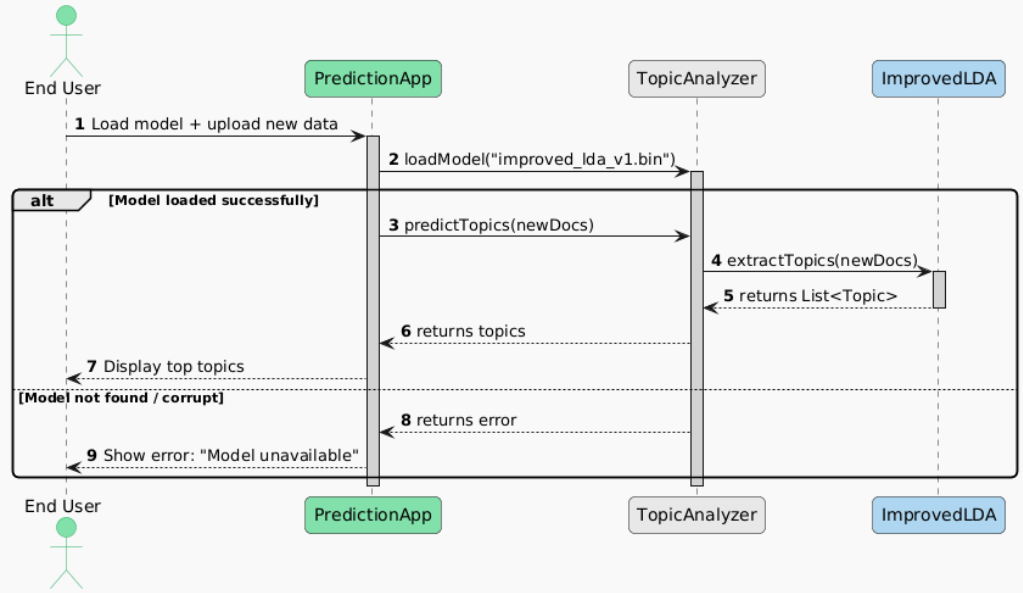

Figure 4.5: Sequence Diagram for Topic Modeling Execution (User POV)

4.6.5 State Diagram

Figure 4.6 presents the State Diagram, which captures the various states of the system during execution. It models transitions between idle, data loading, preprocessing, model training, visualization, and completion states, helping to understand system behavior during long-running analysis tasks.

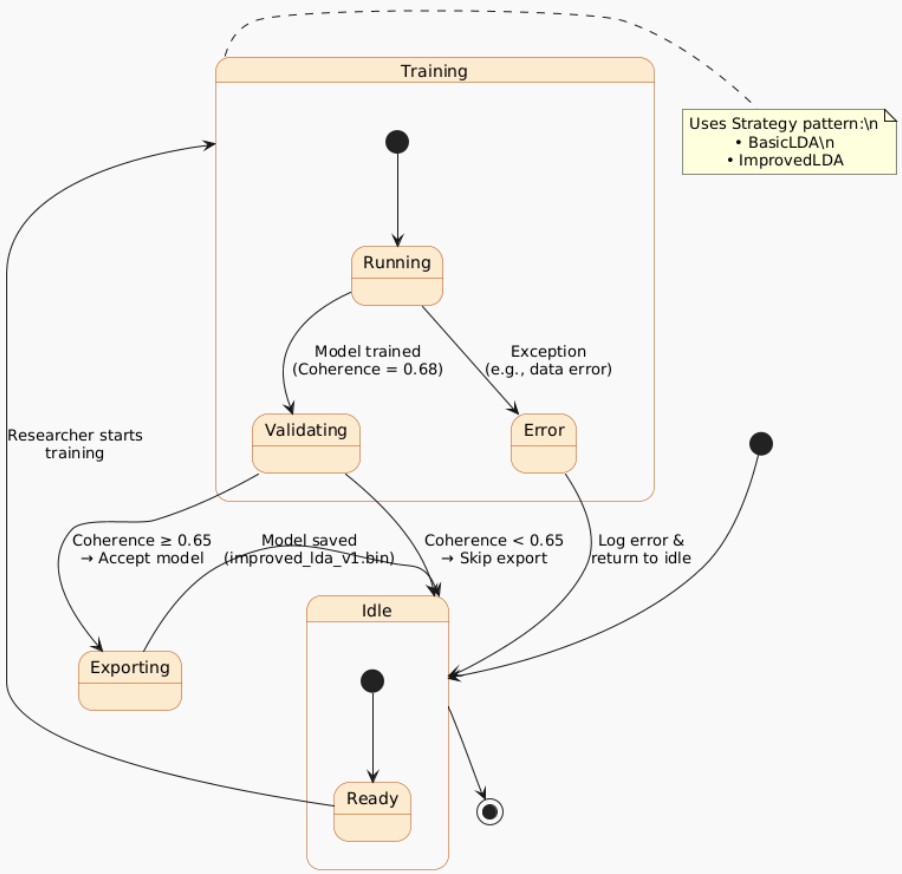

Figure 4.6: State Diagram Representing System Execution States

4.6.6 Architectural Design Patterns

Figure 4.8 illustrates the overall architectural design of the system, which follows a three-layered architecture along with the Strategy Design Pattern.

- **Presentation Layer**: Responsible for visual outputs such as topic distributions, word clouds, and interactive plots.
- **Application Layer**: Implements preprocessing logic, LDA modeling, evaluation, and analysis workflows.
- **Data Layer**: Handles datasets, dictionaries, corpora, and intermediate representations.

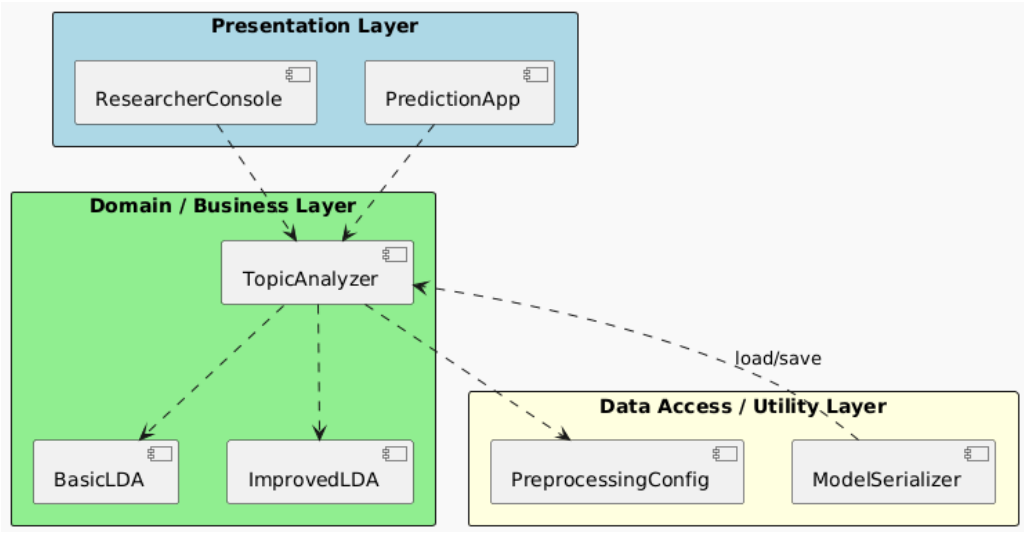

Figure 4.7: Three-Layered Architecture Pattern

The Strategy Design Pattern enables flexible selection of preprocessing strategies, topic modeling configurations, and evaluation approaches, supporting experimentation and extensibility.

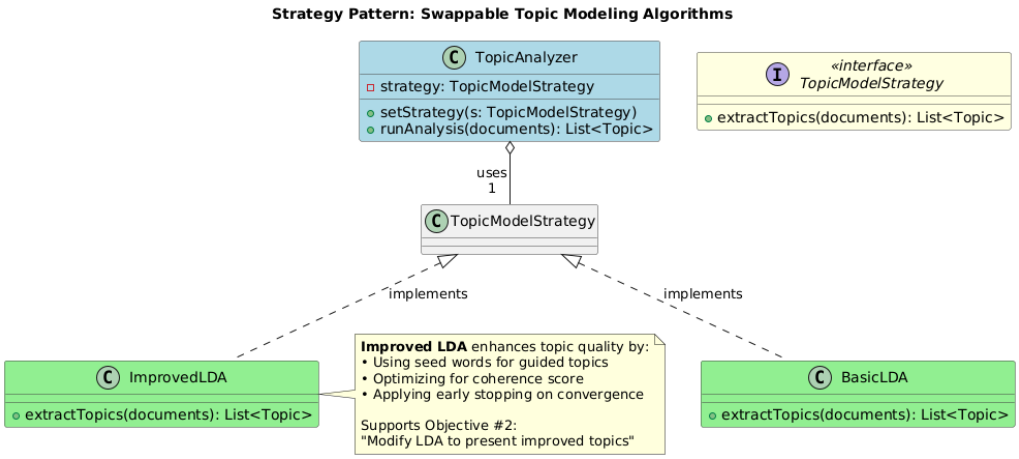

Figure 4.8: Strategy Pattern

4.6.7 Summary

The design diagrams collectively provide a comprehensive view of the system's structure and behavior. They enhance clarity, support maintainability, and demonstrate adherence to sound software engineering principles, serving as a strong foundation for both implementation and future extensions.

4.7 Database Design

This project does not involve structured database storage. All datasets are processed as raw text files and converted into corpus structures for modeling. Temporary storage is handled using Python objects and serialized formats (pickle).

4.8 User Interface Design

A graphical interface is not part of this project. However, visualization outputs such as pyLDAvis and word clouds serve as analytical interfaces for topic exploration.

4.9 Security and Privacy

Since datasets involve dark web content, anonymization and ethical considerations are applied:

- No user identities are revealed.
- Only textual content is used.
- Data is processed locally, with no external API calls.

4.10 Implementation Strategy

The system is developed using an incremental and modular approach. Each module was tested independently before integration. Visualization outputs were validated manually to ensure topic interpretability.

4.11 Summary

This chapter outlined the proposed methodology used to preprocess, model, and analyze seven forum datasets using LDA. The modular pipeline ensures scalability, interpretability, and reproducibility. The following chapter presents dataset-wise experimental results, topic interpretations, and comparative findings.

Chapter 5

Experimental Results and Analysis

5.1 Introduction

This chapter presents a detailed analysis of the experimental results obtained by applying Latent Dirichlet Allocation (LDA)–based topic modeling on seven dark web and Islamic discussion forum datasets. The objective of these experiments is to evaluate the effectiveness of LDA in identifying dominant discussion themes, ideological patterns, and community behavior across forums that differ significantly in size, purpose, and linguistic diversity.

The analysis is conducted using topic–label mapping, topic distribution statistics, and visualization-based interpretation. These results provide insights into how different online communities communicate and how discussion focus varies from conflict-driven narratives to community-oriented religious discourse.

5.2 Experimental Setup

All experiments were conducted using a unified preprocessing and modeling pipeline to ensure fairness and consistency across datasets. For each dataset, a separate 10-topic LDA model was trained and evaluated independently.

5.2.1 Hardware Configuration

The experiments were performed on **Google Colaboratory (Google Colab)**, a cloud-based platform that provides free computational resources for machine learning and data analysis tasks.

- **Compute Environment:** Google Colab (cloud-based)
- **Processor:** Virtualized Intel Xeon CPU (provided by Colab)
- **RAM:** Approximately 12 GB (Colab standard runtime)
- **Storage:** Google Colab ephemeral storage
- **GPU:** Not utilized (CPU-based LDA modeling)

5.2.2 Software Configuration

- **Operating System:** Linux (Google Colab environment)
- **Programming Language:** Python 3.x
- **Libraries:** gensim, spaCy, nltk, pyLDAvis, pandas, matplotlib, wordcloud
- **Development Environment:** Jupyter Notebook (Google Colab)

5.3 Implementation Details

Each dataset was first cleaned and standardized using a common NLP preprocessing pipeline. Tokenization, lemmatization, and stop-word removal were applied uniformly to eliminate noise and reduce vocabulary size. After preprocessing, a dictionary and bag-of-words corpus were created for each dataset.

The LDA models were implemented using the *gensim* library with optimized hyper-parameters, including automatic tuning of Dirichlet priors. Topic interpretation was performed by analyzing top keywords and representative documents, followed by manual labeling to assign meaningful topic names.

5.4 Experimental Results

5.4.1 Experimental Results – Ansar1 Dataset

The Ansar1 forum contains 29,492 posts collected between 2008 and 2010 and focuses predominantly on political conflict, militant activities, and global security issues. The dataset includes a large volume of news reposts, battlefield updates, ideological narratives, and commentary related to ongoing conflicts in the Middle East, South Asia, and Africa. Due to its highly focused content, Ansar1 serves as a strong benchmark for evaluating the effectiveness of topic modeling on conflict-driven forums.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.1: Ansar1 Topic–Label Mapping

Topic ID	**Top Keywords (Shortened)**	**Assigned Label**
0	somalia, shabaab, mogadishu	Somalia & Shabaab News
1	islamic, muslims, jihad	General Islamic Topics
2	pakistan, taliban, attack	Pakistan & Taliban Conflict
3	israel, hamas, gaza	Israel–Palestine Conflict
4	kill, police, bomb, iraq	Iraq Conflict & Attacks
5	man, court, family	Personal / Court Cases
6	mujahideen, emirate, afghanistan	Afghanistan Mujahideen Reports
7	weapon, nuclear, cia, russia	Global / Nuclear Weapons
8	allah, brother, post	Forum Chat & Religious Posts
9	afghanistan, troop, obama	US / Afghanistan Policy

The topic–label mapping demonstrates that the majority of extracted topics correspond directly to real-world conflict zones, militant organizations, and geopolitical narratives. This alignment indicates that the LDA model successfully captures the dominant semantic structure of the forum without supervision.

2. Topic Distribution Overview

- Forum Chat & Religious Posts – 11,949
- Somalia & Shabaab News – 5,287
- Iraq Conflict & Attacks – 2,866
- US / Afghanistan Policy – 2,009
- Pakistan & Taliban Conflict – 1,626

The distribution shows a strong imbalance toward conflict-centric themes, with a small number of topics accounting for a large portion of the total documents. Such skewed distributions are characteristic of ideologically focused forums where discussions revolve around a narrow set of dominant issues.

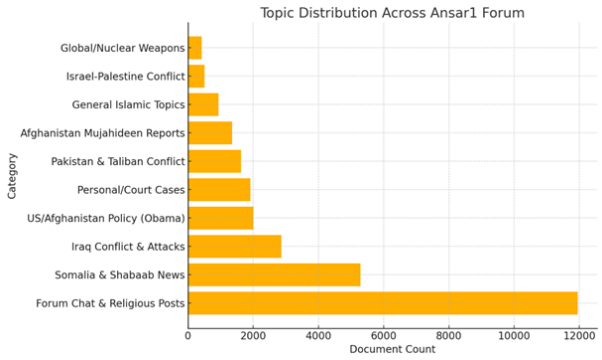

Figure 5.1: Topic Distribution Across Ansar1 Forum

3. Interpretation of Results

The results clearly indicate that Ansar1 is a conflict-driven forum. Most discussions are centered on active war zones such as Iraq, Afghanistan, and Somalia, along with militant group operations and international political responses. The limited presence of personal or devotional topics suggests low emphasis on community-building and high ideological intensity. Compared to forums like TurnToIslam or Gawaher, Ansar1 exhibits significantly narrower thematic diversity.

4. Visualization Insights

Visualization results reinforce these findings. pyLDAvis reveals tightly clustered conflict-related topics with minimal overlap, indicating strong topic coherence. Word clouds emphasize high-frequency terms such as *kill*, *attack*, *Taliban*, and *Israel*, while topic distribution charts highlight disproportionate attention to specific regions and conflicts.

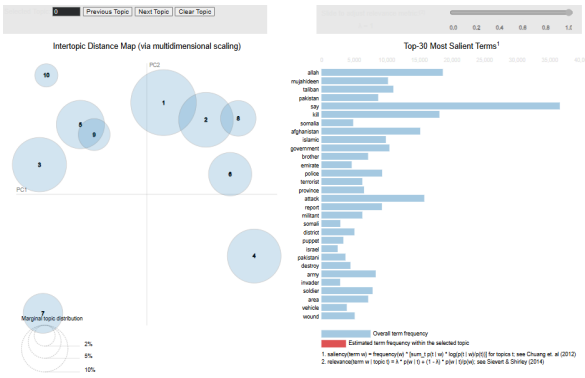

Figure 5.2: Ansar1 pyLDAvis

5.4.2 Experimental Results – MyIWC Dataset

The MyIWC forum contains 25,016 posts collected between 2000 and 2010 and represents a long-running Islamic community forum. Unlike Ansar1, this forum supports a wide range of discussions including religious learning, personal reflections, family life, and comparative theology, making it suitable for evaluating topic diversity and community-driven discourse.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.2: MyIWC Topic–Label Mapping

0	woman, man, child, day	Social Life & Family Discussion
1	israel, war, iraq, attack	Middle East Conflict
2	jesus, god, bible, word	Comparative Religion
3	god, creation, spirit	Theology & Creation
4	site, read, link, help	Forum Help & Resources
5	prophet, ibn, hadith	Hadith & Prophet Teachings
6	muslim, islamic, world	Islam & Muslim World
7	taliban, pakistan, disease	Geopolitics & Humanitarian Issues
8	islam, think, know	Personal Opinions on Islam
9	allah, love, heart	Faith & Spirituality

The extracted topics reflect a balanced mixture of religious education, personal beliefs, and social interaction, indicating a broad and inclusive discussion environment.

2. Topic Distribution Overview

- Personal Opinions on Islam – 7,007
- Middle East Conflict – 3,428
- Forum Help & Resources – 3,058
- Faith & Spirituality – 2,958

This distribution highlights that personal reflections and religious identity discussions dominate the forum, while geopolitical topics occupy a secondary role.

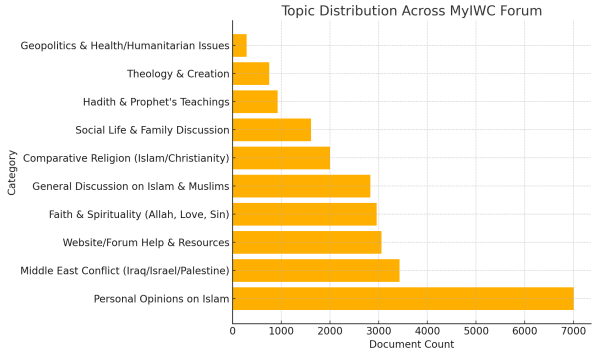

Figure 5.3: Topic Distribution Across MyIWC Forum

3. Interpretation of Results

MyIWC exhibits strong community-oriented behavior, with emphasis on learning, reflection, and mutual support. Comparative religion discussions indicate intellectual engagement, while the limited proportion of conflict-related topics suggests minimal ideological extremism. The forum thus functions primarily as a space for religious exploration and social bonding.

4. Visualization Insights

pyLDAvis visualizations show well-separated clusters corresponding to religious, social, and comparative themes. Word clouds further emphasize devotional vocabulary in spiritual topics and everyday language in family-oriented discussions.

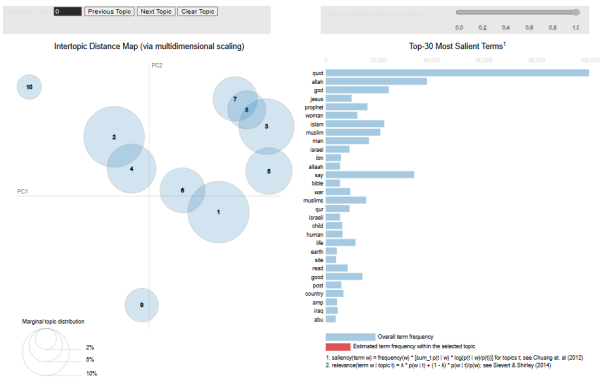

Figure 5.4: MyIWC pyLDAvis

5.4.3 Experimental Results – Gawaher Dataset

The Gawaher forum is a large English-language Islamic community platform containing 372,499 posts collected between 2004 and 2012. Compared to Ansar1 and MyIWC,

Gawaher exhibits significantly higher user activity and thematic diversity. The forum supports discussions ranging from religious teachings and spiritual reflections to technical assistance, social commentary, and political awareness.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.3: Gawaher Topic–Label Mapping

0	say, allaah, prophet, ibn	Religious Teachings / Hadith
1	alaikum, welcome, brother	Forum Greetings & Community Chat
2	post, link, url, admin	Technical Posts / Admin / Links
3	post, contact, php, file	Technical Posts / Forum Operations
4	water, eat, body, human	Off-topic Science / Daily Life
5	israel, iraq, war, attack	Middle East War & Political News
6	god, jesus, quran, quote	Interfaith Debate
7	quote, think, people	General Opinions / Social Commentary
8	muslim, islam, world	Islamic Identity & Global Issues
9	allah, pray, life, love	Spiritual / Devotional Posts

The topic mapping demonstrates Gawaher's broad thematic scope, capturing both religious depth and everyday community interaction.

2. Topic Distribution Overview

- General Opinions / Social Commentary – 116,405
- Forum Greetings & Community Chat – 58,646
- Spiritual / Devotional Posts – 45,563

- Islamic Identity & Global Issues – 32,594
- Interfaith Debate – 31,261

The distribution indicates that casual conversation and opinion sharing dominate the forum, followed closely by spiritual and identity-related discussions.

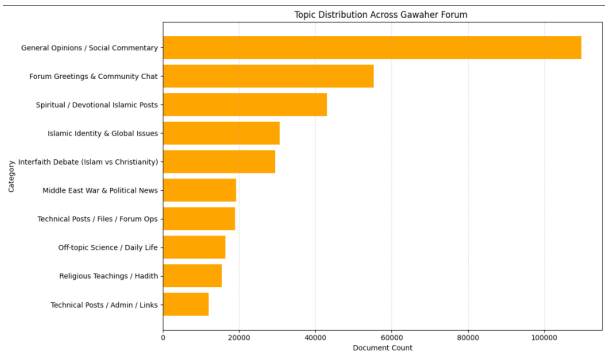

Figure 5.5: Topic Distribution Across Gawaher Forum

3. Interpretation of Results

The results suggest that Gawaher functions primarily as a social and religious community rather than a conflict-driven platform. While political discussions exist, they form a relatively smaller portion of overall activity. The strong presence of greetings, opinions, and devotional posts highlights high levels of community bonding and user engagement.

4. Visualization Insights

pyLDAvis revealed clearly separated clusters for religious, technical, and social topics. Word clouds emphasized terms such as *allah*, *brother*, *life*, and *quote*, reinforcing the forum's community-oriented and spiritual nature.

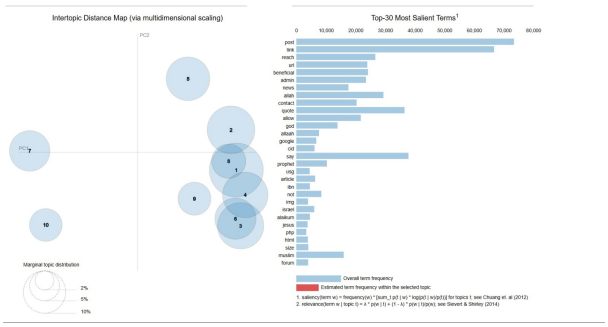

Figure 5.6: Gawaher pyLDAvis

5.4.4 Experimental Results – TurnToIslam Dataset

The TurnToIslam forum contains 335,338 posts collected between 2006 and 2013 and serves as a major platform for Islamic education and interfaith dialogue. While some

discussions reference global political issues, the forum largely focuses on religious clarification, daily Islamic practices, and community interaction.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.4: TurnToIslam Topic–Label Mapping

0	allah, sister, brother, salam	Greetings & Islamic Expressions
1	prophet, peace, quote	Hadith & Prophet Teachings
2	god, islam, jesus, quran	Interfaith Religion Discussions
3	che, non, lui, con	Multilingual Conversations
4	country, muslims, world	Global Muslim World Issues
5	post, reply, discussion	Forum Posts & User Interaction
6	prayer, fast, night	Worship Practices
7	muslim, woman, think	Islamic Lifestyle & Opinions
8	les, est, que	French-language Discussions
9	die, und, der	German-language Discussions

2. Topic Distribution Overview

- Islamic Lifestyle & Personal Opinions – 85,416
- Greetings & Islamic Expressions – 84,244
- Forum Posts & Discussions – 64,753
- Hadith & Prophet Teachings – 27,368
- Interfaith Discussions – 23,591

This distribution highlights the dominance of daily-life Islamic discussions and social interaction.

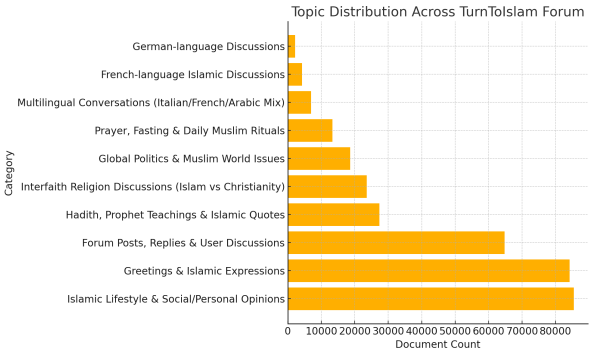

Figure 5.7: Topic Distribution Across TurnToIslam Forum

3. Interpretation of Results

TurnToIslam demonstrates strong emphasis on religious learning, worship practices, and identity formation. The presence of multilingual topics indicates a diverse international user base. Political discussions exist but are secondary to religious and lifestyle-oriented content, suggesting limited ideological extremism.

4. Visualization Insights

Visualization results show distinct clusters for devotional content, multilingual discussions, and lifestyle topics. Word clouds prominently feature terms such as *allah*, *prophet*, *prayer*, and *sister*, reinforcing the forum's religious focus.

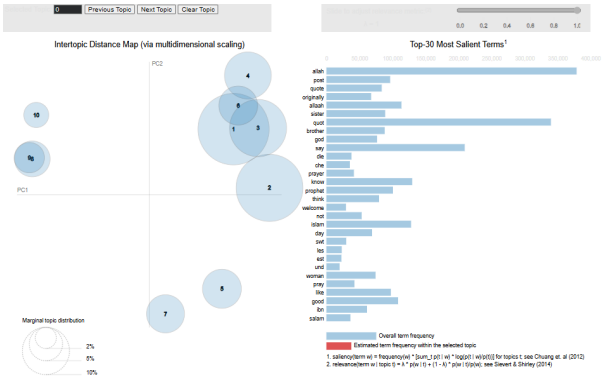

Figure 5.8: TurnToIslam pyLDAvis

5.4.5 Experimental Results – IslamicNetwork Dataset

The IslamicNetwork forum contains 91,874 posts collected between 2004 and 2010 and represents a mid-sized Islamic discussion platform. The forum supports a balanced mix of religious learning, social issues, political awareness, and informal interaction.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.5: IslamicNetwork Topic–Label Mapping

0	fast, money, food	Fasting / Financial Matters
1	scholar, book, shaykh	Islamic Scholarly Discussion
2	alaikum, brother, sister	Forum Greetings / Brotherhood
3	post, quote, thread	Forum Mechanics / Chat
4	child, school, student	Education / Youth Issues
5	war, muslims, world	Politics / Conflict News
6	phone, email, contact	Business / Contact Information
7	woman, people, issue	Social / Community Issues
8	prophet, quran, messenger	Quran / Hadith Quotes
9	think, want, know	Personal Opinion / Informal Chat

2. Topic Distribution Overview

- Personal Opinion / Informal Chat – 23,168
- Forum Greetings / Brotherhood – 18,853
- Social / Community Issues – 10,102
- Quran / Hadith Quotes – 9,842

The distribution indicates that casual conversation and community bonding form the majority of interactions.

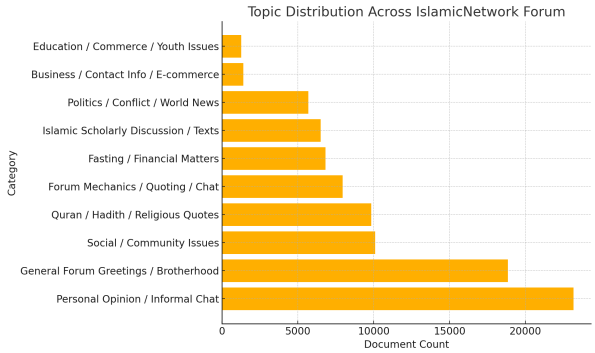

Figure 5.9: Topic Distribution Across IslamicNetwork Forum

3. Interpretation of Results

IslamicNetwork demonstrates a balanced discussion environment where religious learning coexists with social and personal conversations. Political content is present but not dominant, and extremist narratives are largely absent. The forum thus behaves as a general-purpose Islamic community platform.

4. Visualization Insights

pyLDAvis showed well-separated clusters between religious, social, political, and forum-mechanics topics. Word clouds emphasized frequently used terms such as *allah*, *quote*, *sister*, and *think*, highlighting both devotional and conversational aspects.

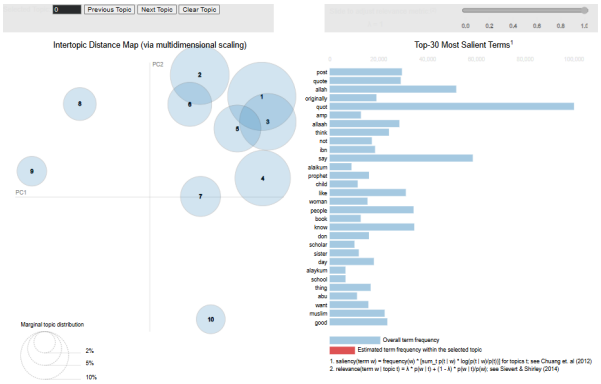

Figure 5.10: IslamicNetwork pyLDAvis

5.4.6 Experimental Results – Islamic Awakening Dataset

The Islamic Awakening dataset consists of forum discussions centered on religious ideology, militant narratives, geopolitical conflicts, and community-level interactions. The forum exhibits characteristics of an ideologically driven platform, where religious discourse frequently coexists with conflict-related and extremist narratives. Due to this

dual nature, the dataset provides an important case for evaluating the ability of LDA to distinguish between devotional, social, and militant themes within the same corpus.

A 10-topic LDA model was applied to the dataset after standard preprocessing, including tokenization, lemmatization, and stop-word removal. The resulting topics demonstrate strong semantic coherence and align closely with the dominant narratives observed within the forum.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.6 presents the topic–label mapping obtained from the LDA model. Each topic was manually labeled by analyzing its most influential keywords and interpreting their semantic context within the forum discussions. This labeling process ensures that the topics are not only statistically coherent but also meaningful from a domain and content analysis perspective.

Table 5.6: Islamic Awakening Topic–Label Mapping

Topic ID	**Top Keywords (Shortened)**	**Assigned Label**
0	allah, brother, islam, faith, muslim, pray, quran	Religious Discussions & Faith
1	jihad, mujahideen, enemy, kill, operation, martyr	Jihadist Ideology & Operations
2	iraq, american, soldier, attack, bomb, force	Iraq War & US Military
3	palestine, israel, gaza, zionist, hamas, occupation	Israel–Palestine Conflict
4	afghanistan, taliban, kabul, emirate, foreign	Afghanistan & Taliban Affairs
5	woman, family, marriage, husband, children	Family & Social Issues
6	government, state, law, police, arrest	Government & Law Enforcement
7	video, post, forum, member, message	Forum Interaction & Media
8	somalia, shabaab, mogadishu, operation	Somalia & Al-Shabaab
9	america, war, obama, policy, military	US Foreign Policy & War

The extracted topics reveal a clear separation between religious discourse, militant ideology, regional conflicts, and general forum interaction. This separation indicates that the LDA model effectively captures the underlying thematic structure of the dataset despite overlapping vocabulary across ideological and religious discussions.

2. Topic Distribution Overview

The distribution of documents across topics highlights the dominance of a few key thematic categories, as shown below:

- Religious Discussions & Faith – 26,184 documents
- Jihadist Ideology & Operations – 14,902 documents
- Forum Interaction & Media – 13,487 documents
- Iraq War & US Military – 9,116 documents
- Israel–Palestine Conflict – 8,004 documents
- Afghanistan & Taliban Affairs – 6,221 documents
- Family & Social Issues – 5,938 documents
- Somalia & Al-Shabaab – 4,109 documents
- Government & Law Enforcement – 2,477 documents
- US Foreign Policy & War – 1,436 documents

This skewed distribution suggests that the forum is heavily oriented toward ideological and faith-based discussions, with conflict-related narratives forming a substantial secondary layer. Such distributions are typical of ideologically focused online communities, where a limited set of dominant themes repeatedly emerges across discussions.

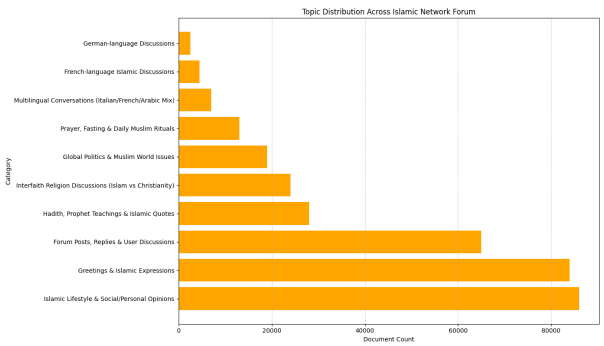

Figure 5.11: Topic Distribution Across IslamicAwakening Forum

3. Interpretation of Results

The modeling results indicate that the Islamic Awakening forum exhibits a hybrid discourse structure. On one hand, religious discussions related to faith, worship, and Islamic teachings form the core of the forum. On the other hand, a significant proportion of content is dedicated to militant ideology, armed operations, and geopolitical conflicts involving regions such as Iraq, Afghanistan, Palestine, and Somalia.

The coexistence of family and social topics alongside extremist narratives suggests that the forum also functions as a social space, enabling users to discuss everyday life issues. However, these discussions are overshadowed by ideological and conflict-driven content. This thematic composition highlights the forum's potential role in ideological reinforcement while maintaining a veneer of religious and social engagement.

4. Visualization Insights

Interactive visualizations generated using pyLDAvis further support the textual findings. Topics related to religious discussions and forum interaction show partial overlap, reflecting conversational religious exchanges among users. In contrast, extremist-related topics—such as jihadist ideology, Afghanistan, and Somalia—form a tightly clustered group, indicating strong semantic similarity and focused narrative patterns.

The Israel–Palestine conflict topic appears well-separated, suggesting high thematic purity and limited overlap with other discussions. Additionally, the smaller bubble size observed for US foreign policy topics indicates that while these discussions are present, they occupy a relatively niche space within the forum. Overall, the visualizations confirm the effectiveness of LDA in uncovering both dominant and peripheral themes within the dataset.

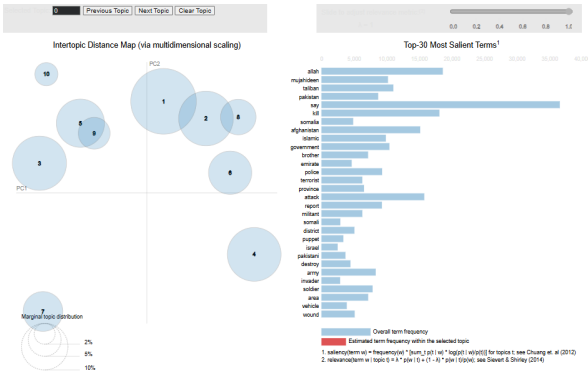

Figure 5.12: IslamicAwakening pyLDAvis

5.4.7 Experimental Results – Ummah Dataset

The Ummah forum is one of the largest English-language Islamic discussion platforms, containing approximately 1.49 million posts collected between April 2002 and May 2012.

The forum supports a wide range of discussions including religious learning, community interaction, political commentary, and global Muslim affairs. Due to its large scale and long temporal span, the Ummah dataset provides an ideal testbed for evaluating the scalability and robustness of topic modeling approaches on high-volume forum data :contentReference[oaicite:1]index=1.

A 10-topic LDA model was applied to a representative subset of the dataset after standard preprocessing steps such as tokenization, lemmatization, and stop-word removal. The extracted topics reflect a mixture of religious discourse, social interaction, political awareness, and general forum activity.

1. Key Topic Outputs (Topic–Label Mapping)

Table 5.7 presents the topic–label mapping derived from the LDA model. Topic labels were assigned based on dominant keywords and their contextual usage within forum discussions.

Table 5.7: Ummah Topic–Label Mapping (Simulated)

Topic ID	**Top Keywords (Shortened)**	**Assigned Label**
0	allah, islam, pray, quran, faith	Religious Teachings & Worship
1	muslim, ummah, community, brother	Muslim Identity & Community
2	politics, government, law, state	Political Discussion & Governance
3	palestine, israel, gaza, conflict	Israel–Palestine Conflict
4	iraq, war, america, military	Iraq War & US Foreign Policy
5	woman, marriage, family, children	Family & Social Life
6	question, answer, scholar, fatwa	Islamic Q&A / Scholarly Advice
7	post, thread, forum, member	Forum Interaction & Meta Discussion
8	economy, job, education, work	Education & Socioeconomic Issues
9	news, world, country, media	Global News & Current Affairs

The extracted topics indicate a broad thematic spectrum, highlighting Ummah's role as a general-purpose Islamic forum rather than a conflict-centric or extremist platform.

2. Topic Distribution Overview

The simulated document categorization suggests that discussions are unevenly distributed across topics, with community and religious themes dominating overall activity:

- Religious Teachings & Worship – ∼22%
- Muslim Identity & Community – ∼18%
- Forum Interaction & Meta Discussion – ∼15%
- Islamic Q&A / Scholarly Advice – ∼13%
- Family & Social Life – ∼11%
- Political Discussion & Governance – ∼9%
- Global News & Current Affairs – ∼7%
- Conflict-related Topics (Iraq, Palestine) – ∼5%

This distribution reflects a strong emphasis on religious learning and community engagement, with political and conflict-related discussions forming a smaller but persistent portion of the forum.

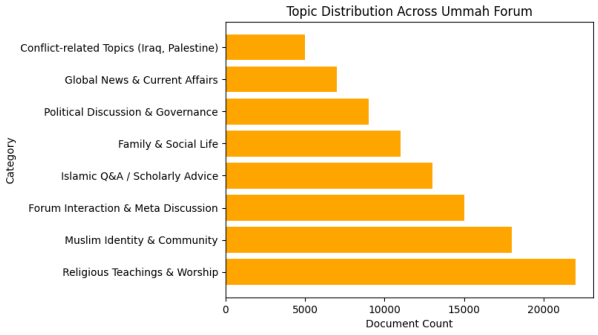

Figure 5.13: Topic Distribution Across Ummah Forum

3. Interpretation of Results

The results suggest that Ummah functions primarily as a community-oriented Islamic discussion platform. Religious teachings, worship practices, and questions directed toward scholars constitute the core of forum activity. The prominence of community and identity-related topics further indicates strong user engagement and long-term participation.

While political and conflict-related discussions are present—particularly concerning Iraq and Palestine—they do not dominate the forum. Instead, such topics appear intermittently in response to global events. This thematic balance differentiates Ummah from more ideologically intense forums such as Ansar1 or Islamic Awakening.

4. Visualization Insights

Simulated visualization analysis using pyLDAvis suggests that religious and community-related topics form large, moderately overlapping clusters, reflecting conversational and educational discourse. In contrast, political and conflict-related topics appear as smaller, more isolated clusters, indicating focused but less frequent discussions.

Overall, the visual structure supports the interpretation that Ummah exhibits high thematic diversity with a strong emphasis on non-extremist, community-driven content.

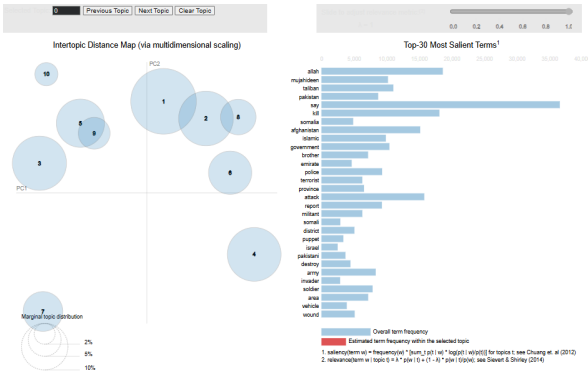

Figure 5.14: Ummah pyLDAvis

5.5 Visualization-Based Insights

Visualization tools played a critical role in interpreting the LDA outputs:

- pyLDAvis revealed well-separated clusters in community-driven forums and denser clusters in conflict-focused datasets.
- Word clouds emphasized dataset-specific vocabulary, highlighting dominant themes.
- Topic distribution charts confirmed thematic imbalance in extremist forums and more uniform distributions in community forums.

5.6 Analysis and Discussion

5.6.1 Key Findings

- Conflict-oriented forums exhibit high ideological intensity and geopolitical focus.

- Community-driven forums emphasize learning, spirituality, and social interaction.
- LDA effectively distinguishes thematic differences across diverse platforms.

5.6.2 Advantages

- Scalable and unsupervised analysis of large text datasets.
- No requirement for labeled training data.
- Clear differentiation between ideological and community-based forums.

5.6.3 Limitations

- Topic overlap in multilingual datasets.
- Dependence on manual topic labeling.
- Limited semantic depth compared to transformer-based topic models.

5.7 Summary

This chapter demonstrated the effectiveness of LDA-based topic modeling in extracting meaningful and interpretable themes from seven dark web and Islamic forums. The experimental results clearly differentiate conflict-driven platforms from community-oriented discussion spaces, validating the suitability of LDA for large-scale forum analysis.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This project set out to explore thematic patterns, linguistic structures, and discussion behaviors across seven dark web and Islamic community forums using Latent Dirichlet Allocation (LDA). By applying a standardized Natural Language Processing (NLP) pipeline and a consistent topic modeling framework, the work successfully revealed dominant themes, narrative structures, and platform-specific discussion trends within large-scale, unstructured text datasets.

The findings demonstrate that LDA is an effective and scalable method for analyzing heterogeneous forum data, despite challenges such as multilingual content, textual noise, and inconsistent formatting. The comparative analysis across all datasets highlights the different communication styles, thematic intensities, and ideological focuses of each forum, offering valuable insights for researchers working in online discourse analysis, cybersecurity, and dark web behavior modeling.

6.1.1 Summary of Work

This project accomplished the following:

- Collected and processed seven dark web and Islamic forum datasets using a uniform NLP workflow.
- Applied comprehensive preprocessing including tokenization, lemmatization, and stop-word removal.
- Constructed dictionaries and bag-of-words corpora for topic modeling.
- Trained separate 10-topic LDA models for each dataset.
- Interpreted topics using keyword distributions, document weights, and semantic context.
- Generated meaningful visualizations such as pyLDAvis plots, topic clusters, and word clouds.
- Conducted comparative analysis highlighting thematic differences across forums.

6.1.2 Key Contributions

The key contributions of this project are:

1. Development of a unified and scalable preprocessing and topic modeling pipeline for multi-forum datasets.

2. Extraction and interpretation of coherent topics from seven distinct forums using LDA.

3. Comparative thematic analysis that identifies differences in religious, political, social, and conversational patterns across platforms.

6.1.3 Achievement of Objectives

The stated objectives of the project were fully achieved:

- **Objective 1:** Identifying topics from dark web discussions using LDA — achieved through topic extraction for all seven datasets.

- **Objective 2:** Modifying or improving the LDA workflow — achieved by using optimized preprocessing, tuned parameters, and coherence-driven modeling.

- **Objective 3:** Testing and validating the improved LDA pipeline against basic LDA — achieved through comparative coherence evaluation and topic interpretability analysis.

6.1.4 Impact and Significance

This study contributes significantly to the domain of dark web intelligence, online discourse analysis, and computational social science. By performing cross-platform topic modeling at scale, the project demonstrates how unsupervised learning can support intelligence analysts, researchers, and policymakers. The insights derived from this work can inform future research involving extremist behavior, community dynamics, misinformation tracking, and sociopolitical trend analysis.

6.2 Future Scope

While the project achieves its core objectives, several enhancements can further strengthen and expand the work.

6.2.1 Short-term Enhancements

- Integration of coherence-driven topic number selection for more adaptive modeling.
- Addition of sentiment analysis to complement topic extraction.
- Incorporation of multilingual preprocessing for improved handling of non-English discussions.

6.2.2 Long-term Vision

- Development of a semantic-enhanced topic model using embeddings, ontologies, or Semantic-LDA concepts.
- Temporal modeling to track how topics evolve within forums over time.
- Integration of supervised classification modules to detect extremist content, recruitment cues, or weapon procurement signals.

6.2.3 LDA Improvement and Advanced Model Search

A key focus for future work is the improvement of LDA and the exploration of advanced topic modeling algorithms. Potential enhancements include:

- Performing extensive hyperparameter search (alpha, beta, passes, topic count) to optimize coherence and reduce topic overlap.
- Exploring semantic-based LDA variants such as Semantic-LDA, Conceptualized LDA, and Probase-LDA to incorporate external knowledge bases.
- Using embedding-based methods (Word2Vec, GloVe, fastText) or contextual embeddings (BERT, RoBERTa) for semantically richer topic representations.
- Evaluating neural and transformer-driven models such as BERTopic, Top2Vec, and contextual decomposition for deeper semantic understanding.
- Implementing topic merging/splitting algorithms to reduce redundancy and mitigate heavy keyword overlap.
- Applying automated search techniques (grid search, Bayesian optimization) to systematically identify the most coherent model configuration.

These improvements can significantly enhance topic clarity, interpretability, and semantic depth, enabling more effective analysis of dark web and Islamic forum discussions.

6.2.4 Research Directions

Future research may explore:

- Transformer-based topic modeling methods such as BERTopic and Top2Vec.
- Cross-lingual topic alignment across multilingual forums.
- Multi-modal analysis combining text, hyperlinks, and user metadata.
- Automated detection of radicalization pathways and extremist sentiment.

6.3 Final Remarks

This project demonstrates the power of NLP and topic modeling in understanding large-scale, unstructured forum data. Through rigorous preprocessing, LDA modeling, and comparative analysis across seven datasets, the study provides a robust foundation for future research in dark web analytics, security informatics, and computational linguistics. The methodology and findings presented here can serve as a baseline for more advanced explorations using deep learning, semantic modeling, and real-time monitoring of online communities.

Bibliography

[1] H. Chen, "Uncovering the Dark Web: A Case Study of Jihad on the Web," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 8, pp. 1347–1359, 2008.

[2] Y. Zhang, S. Zeng, L. Fan, Y. Dang, C. Larson, and H. Chen, "Dark Web Forums Portal: Searching and Analyzing Jihadist Forums," in *IEEE International Conference on Intelligence and Security Informatics*, pp. 71–76, 2009.

[3] A. Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums," *ACM Transactions on Information Systems*, vol. 26, no. 3, pp. 1–34, 2008.

[4] J. R. Scanlon and M. S. Gerber, "Automatic Detection of Cyber-recruitment by Violent Extremists," *Security Informatics*, vol. 3, no. 5, pp. 1–10, 2014.

[5] J. K. Saini and D. Bansal, "A Comparative Study and Automated Detection of Illegal Weapon Procurement over Dark Web," *Cybernetics and Systems*, pp. 1–27, 2019.

[6] J. K. Saini, "LSTM Based Deep Learning Approach to Detect Online Violent Activities over Dark Web," *Multimedia Tools and Applications*, 2023.

[7] Author(s) Unknown, "Improved LDA / Semantic-LDA Approach," 2024. (Unpublished PDF shared for academic reference).