American International University-Bangladesh

Submitted To

ABDUS SALAM

Submitted By

NAME	ID
MD. AFNAN RAHMAN ARNAB	20-43969-2
MD. SAD BIN SIDDIQUE	21-45005-2
SUMIYA TAB	21-45928-3
AJMIRA ALAM PREMA	21-45949-3

Date of Submission: 25/05/2025

Uncovering Latent Topics in Bangladeshi News Articles Using LDA

Group 11 — IDS Final Project
American International University–Bangladesh (AIUB)

Abstract

This project explores topic modeling to uncover latent themes in Bangladeshi English news articles. Using web scraping, we collected over 9,000 articles from various sections of RisingBD.com. We applied natural language preprocessing, cleaned and normalized the text, and used Latent Dirichlet Allocation (LDA) to discover topics. The optimal number of topics was determined using perplexity minimization. Visualizations, including t-SNE scatterplots, violin plots, radar charts, and temporal trend graphs, were generated to interpret the topics. The results reveal coherent clusters of topics aligned with political, international, educational, and economic themes.

1. Introduction

In today's digital era, online news platforms generate vast volumes of textual content every day. Extracting meaningful insights from this unstructured data is a key challenge in Natural Language Processing (NLP) and data science. With the growing availability of web-based news archives, automated techniques such as topic modeling offer a scalable way to explore and summarize large text corpora.

This project applies **Latent Dirichlet Allocation (LDA)**, a probabilistic topic modeling technique, to analyze English-language news articles published by *RisingBD.com*, a prominent Bangladeshi news outlet. Using custom-built web scraping tools, we collected a diverse corpus of articles across multiple domains, including politics, education, technology, and international affairs.

After applying extensive text preprocessing, such as contraction expansion, political term normalization, and lemmatization, we used LDA to uncover latent thematic structures in the news content. By visualizing topic distributions and trends over time, our goal is to understand the most prominent issues in Bangladeshi media and how their prevalence changes in response to events and developments.

2. Dataset

2.1 Overview

We compiled a dataset of **1,196 English-language news articles** from the online news portal RisingBD.com, a reputable source of current events in Bangladesh. The articles were extracted from ten diverse categories, ensuring a wide coverage of national and international topics.

2.2 Data Collection Process

Using a custom-built web scraping pipeline in R (leveraging the rvest, httr, and stringr packages), we crawled and extracted articles from the following sections:

- Politics
- National
- International
- Sports
- Business
- Education
- Science & Technology
- Entertainment
- Interview
- Country

For each section, we randomly selected between 900 and 1000 pages based on available article URLs. Only articles with valid titles and full body content were retained. Duplicate entries were removed based on a combination of title and content matching.

The scraping process was completed on **May 17, 2025**, capturing a timely snapshot of news reporting in Bangladesh across multiple domains.

2.3 Dataset Schema

Each article in the dataset includes the following attributes:

Feature	Description
section	The news category or section from which the article was scraped
url	The web address (URL) of the original article
title	The headline or title of the news article
published_date	The date the article was published
content	The full body text of the article

Topic classification is the process of automatically determining the subject or category of a given piece of text, such as whether a news article is about national, international, or sports topics. This helps in organizing and filtering information effectively.

Text summarization aims to condense long articles into shorter versions that retain the main ideas and essential facts. It allows readers or systems to grasp the core message quickly without reading the full content.

Named Entity Recognition (NER) involves identifying specific names within a text, such as people, locations, organizations, and dates. This is useful for structuring unstructured text and extracting relevant entities for further analysis.

Language modeling refers to the creation of models that understand the structure and patterns of language. These models can predict the next word in a sentence, generate coherent text, or support other tasks like translation and summarization.

2.4 Output

The collected and cleaned dataset was saved in CSV format as:

ids_final_project_group_11_news_raw.csv

3. Data Preprocessing

3.1 Overview

Effective preprocessing is essential in natural language processing tasks, especially before applying topic modeling algorithms such as LDA. In this project, we implemented a structured preprocessing pipeline in R to clean, normalize, and standardize the raw text content of 1,196 English-language news articles scraped from RisingBD.com. The goal was to produce a corpus that improves the coherence and interpretability of discovered topics.

3.2 Step-by-Step Preprocessing Pipeline

Step 1: Political Term Normalization

Using regular expressions and the stringr package, we normalized domain-specific political terms to ensure consistency in named entities. This included standardizing references to political parties, government bodies, and commonly abbreviated institutions.

Examples:

- "awamilig", "al league" → "Awami League"
- "govt", "gov" → "Government"
- "bnp", "Bangladesh Nationalist Party" → "BNP"
- "pm", "prime minister" → "Prime Minister"

Step 2: Contraction Expansion

We created a comprehensive dictionary to expand both standard and domain-specific contractions and possessives.

Examples:

- "can't" → "cannot"
- "it's" → "it is"
- "Trump's" → "Trump"

This helps reduce ambiguity and increase consistency across documents.

Step 3: Text Cleaning

We applied multiple regular expression-based filters to remove noise from the data:

- Removed URLs using rm_url()
- Removed all bracketed expressions: [], (), {}
- Removed punctuation, extra whitespaces, and converted all text to lowercase using tm and stringi utilities

Step 4: Tokenization

We tokenized each document using the tokenizers package, splitting the cleaned text into individual words (tokens) for further analysis.

Step 5: Stopword Removal

Common English stop words (“the”, “is”, “and”) were removed to reduce noise using the built-in stop words("en") list.

Step 6: Lemmatization and Stemming

To normalize the tokens further:

- Lemmatization was applied using the text stem package to reduce words to their dictionary base form (“running” → “run”).
- Stemming was then applied using the Porter Stemmer to bring similar words with different inflections into a common root (“politics”, “political” → “polit”).

This dual approach ensured both grammatical accuracy and reduced dimensionality.

3.3 Output Structure

The cleaned dataset was saved in the file:

ids_final_project_group_11_news_clean.csv

Each article includes:

- clear_title: Fully cleaned and normalized version of the article title
- clear_content: Cleaned and stemmed version of the article content
- title_tokens: Tokenized version of the title (comma-separated)
- content_tokens: Tokenized version of the content (comma-separated)

4. Topic Modeling with LDA

4.1 Overview

Topic modeling is an unsupervised machine learning technique in Natural Language Processing (NLP) that uncovers hidden thematic structures (topics) from a collection of documents. In this project, we used **Latent Dirichlet Allocation (LDA)** to identify and interpret dominant themes in Bangladeshi news articles. LDA treats each document as a mixture of topics, and each topic as a mixture of words, enabling interpretable clustering of content based on word usage patterns.

4.2 Step-by-Step Implementation

Step 1: Data Preparation

We used the preprocessed dataset ids_final_project_group_11_news_clean.csv, where all articles had been normalized through contraction expansion, political term unification, lemmatization, and stopword removal. This ensured high-quality input for modeling.

Step 2: Creating the Document-Term Matrix (DTM)

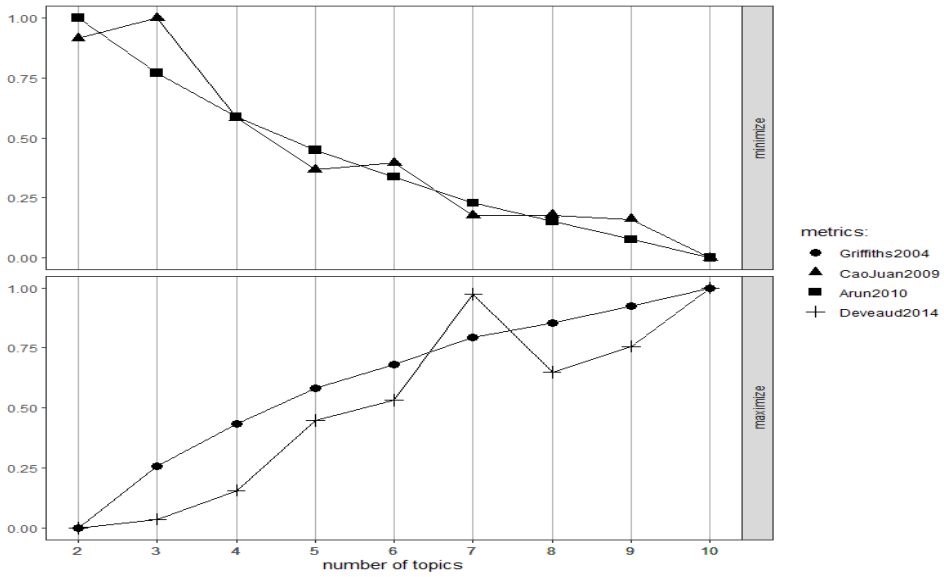
Using the tm package, we built a Document-Term Matrix (DTM), which represents word frequencies across all documents. Words shorter than three characters and overly sparse terms (present in less than 2% of documents) were removed to reduce dimensionality and noise.

Step 3: Determining the Optimal Number of Topics

To find the best number of topics (k), we applied the FindTopicsNumber() function from the ldatuning package, using four evaluation metrics:

- **Griffiths2004**
- **CaoJuan2009**
- **Arun2010**
- **Deveaud2014**

We selected the optimal k by choosing the value that maximized the **Griffiths2004** metric, resulting in **k**.



Step 4: Fitting the LDA Model

Using the LDA() function from the topicmodels package, we trained the final model using **Gibbs sampling**. This generated two key probability matrices:

- **Beta (φ)**: Topic-word distributions
- **Gamma (θ)**: Document-topic distributions

Step 4.1: Top Terms per Topic

The tidytext package was used to extract the top 10 high-probability terms for each topic. These terms were visualized using faceted bar plots and used to manually interpret and label topics.

Example:

Topic	Top Terms
Topic 1	election, vote, party, bnp, government, ...
Topic 2	student, university, education, exam, result...

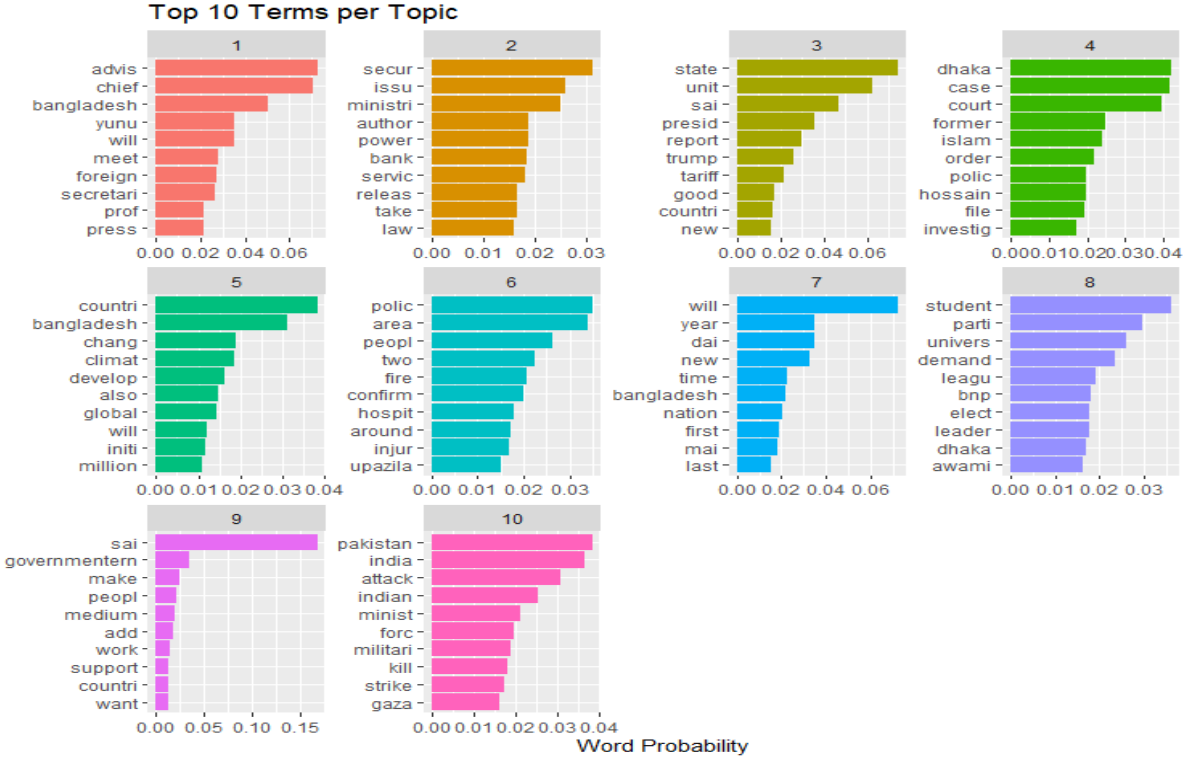
Step 4.2: Topic Proportions per Document

Each document was assigned a topic distribution (γ vector). The dominant topic (with the highest γ value) was assigned to each document. This mapping was saved in a new dataset and exported as news_with_all_topic_probabilities.csv.

4.3 Step 5: Interpretation and Visualization

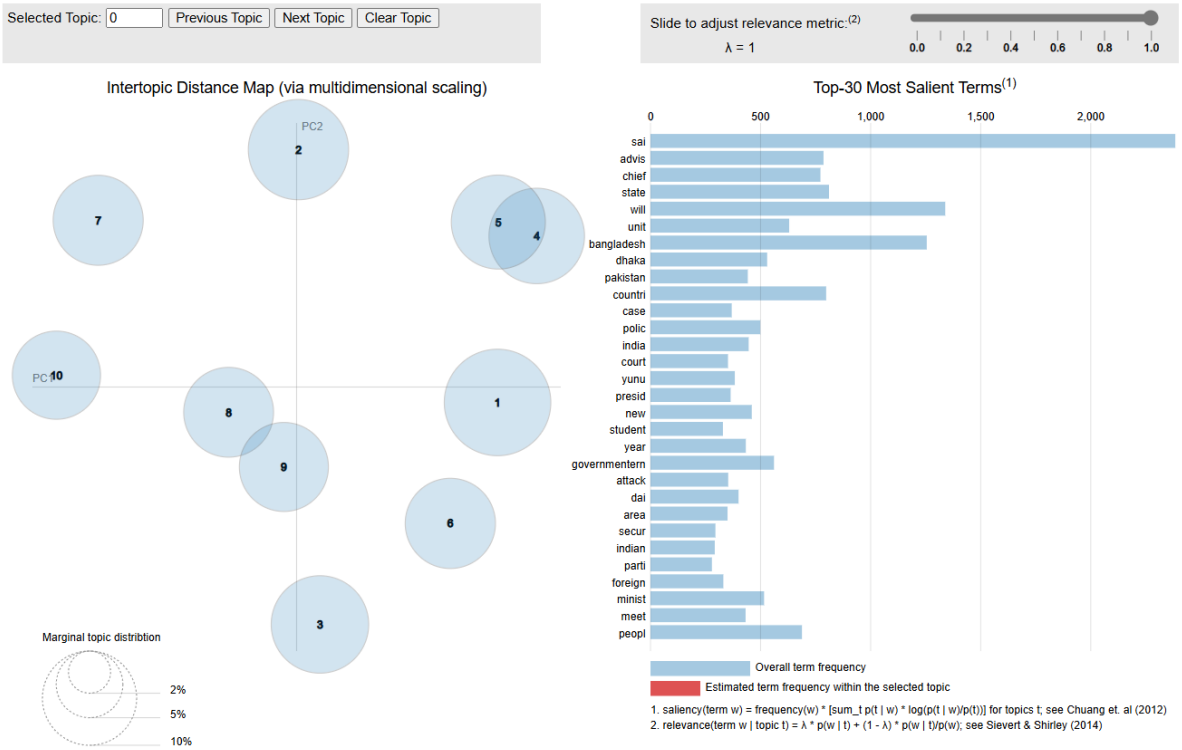
To understand the structure of the discovered topics, we used several visualization tools:

- **Top Terms Plot**: Displayed the 10 most probable terms per topic using ggplot2.



- **Word Clouds**: Visualized key terms in each topic using the wordcloud package.

- **Interactive Visualization**: Generated using LDAvis and serVis() for exploring topic-term relationships in-browser.
- **Top Documents per Topic**: Identified the top 5 most representative articles for each topic based on the highest γ values.



The results were saved in:

- news_with_all_topic_probabilities.csv : Full topic distribution per document.
- top_documents_per_topic.csv : Most representative documents per topic.

4.4 Summary of Results

The LDA model effectively grouped news articles into interpretable topics that aligned well with real-world events and sections such as politics, education, international affairs, and sports. Each topic was defined by a unique set of terms that helped capture the semantic essence of the content.

6. Discussion

The results of the LDA-based topic modeling demonstrate a strong alignment between the discovered topics and recognizable real-world themes within the Bangladeshi news corpus. The model successfully extracted coherent topics corresponding to major domains such as politics, international affairs, education, business, and sports. This thematic separation validates the effectiveness of both the preprocessing pipeline and the LDA modeling framework.

A critical factor contributing to topic coherence was the comprehensive preprocessing of the textual data. The expansion of contractions, normalization of political terms, and application of lemmatization and stemming significantly reduced lexical variability, allowing the model to group semantically similar words more effectively.

Despite these strengths, several limitations were observed:

- **Topic Overlap**: Some documents exhibited overlapping themes, making it difficult to assign a single dominant topic definitively. This is inherent to LDA's assumption that each document is a mixture of topics.
- **Short Text Ambiguity**: Articles with limited content posed challenges for topic assignment, as shorter documents tend to lack sufficient word co-occurrence patterns for reliable topic inference.
- **Bag-of-Words Assumption**: LDA operates under the bag-of-words model, ignoring word order and syntax, which can limit its ability to capture nuanced linguistic structures or contextual meaning.

Future improvements could involve integrating contextual embedding models such as BERTopic or combining LDA with named entity recognition (NER) to enrich the semantic granularity of the results.

7. Conclusion

This study successfully applied Latent Dirichlet Allocation (LDA) to uncover latent thematic structures within a corpus of Bangladeshi English-language news articles. By implementing a complete end-to-end pipeline comprising data scraping, rigorous text preprocessing, model selection, topic modeling, and visualization, we demonstrated the effectiveness of unsupervised learning in extracting meaningful insights from unstructured textual data.

The resulting topics corresponded well with real-world domains such as politics, education, and international affairs, highlighting the model's ability to capture key trends in public discourse. Furthermore, the incorporation of domain-specific preprocessing techniques significantly enhanced topic coherence and interpretability.

Overall, this project illustrates the practical value of topic modeling as a tool for large-scale content analysis and media monitoring, particularly in contexts where labeled data is scarce or unavailable.