American International University-Bangladesh

FACULTY OF SCIENCE & TECHNOLOGY

INTRODUCTION TO DATA SCIENCE

Spring 2024-25

Section: B

Group 10

Supervised by

TOHEDUL ISLAM

Assistant Professor, Computer Science

Submitted by:

Name	***ID***
Md Tareq Jamil Sarkar	**22-46619-1**
Md Mehedi Hasan	**22-46322-1**
Hasin Almas Sifat	**22-48679-3**

Submitted On: 23 june 2025

Introduction

First, we scraped online news articles related to Middle East issues from the Prothom Alo English website. The objective is to gather unstructured text data from multiple webpages, clean and preprocess the textual content, and apply topic modeling using Latent Dirichlet Allocation (LDA). By identifying latent themes from the articles, we aim to explore topic distributions and term relevance across documents.

Dataset Description

The data consists of 10 different English news articles collected from:

https://en.prothomalo.com/international

Each article includes the title, content paragraphs, and meta structure. After scraping, the texts were aggregated and structured into a CSV file containing the title and corresponding paragraph texts.

Required Libraries

```
install.packages("rvest")
install.packages("tm")
install.packages("SnowballC")
install.packages("tidytext")
install.packages("dplyr")
install.packages("textstem")
install.packages("topicmodels")
install.packages("textmineR")
install.packages("LDAvis")
install.packages("ggplot2")
install.packages("slam")
install.packages("wordcloud")
```

These libraries were used for:

- Scraping (rvest)
- Text mining & cleaning (tm, tidytext, textstem)
- Topic modeling (topicmodels)
- Visualization (ggplot2, wordcloud, LDAvis)

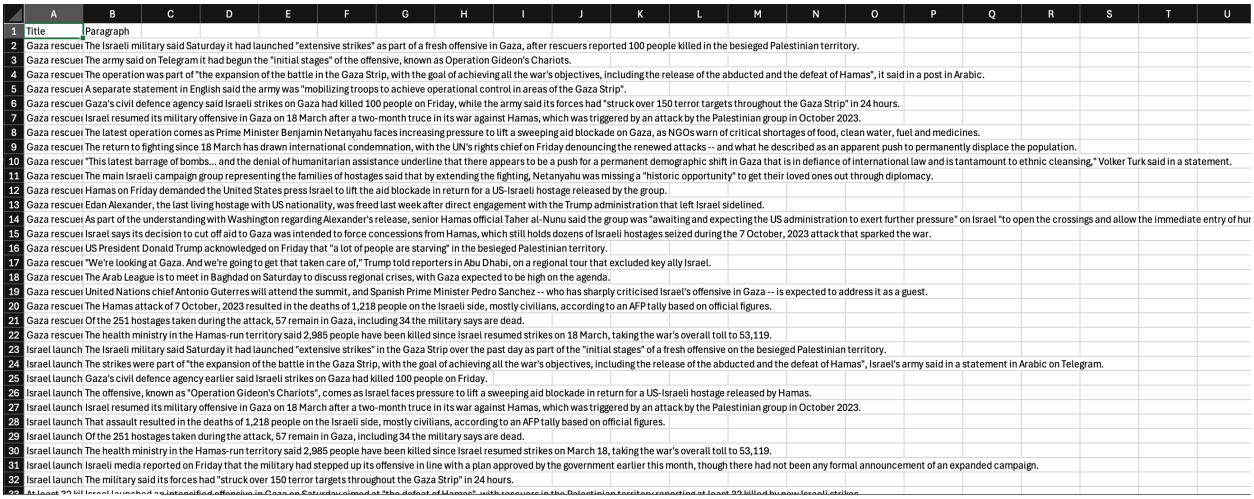
Web Scraping

Ten separate news articles were scraped using rvest. HTML tags such as <title> and <p> were extracted and converted into text using:

```
html_node(link, 'p') %>% html_text()
```

The title and paragraphs were combined into data frames and merged using rbind() to form a complete dataset. The final result was exported into:

```
write.csv(scraped_data, "scraped_data.csv")
```



	Title	Paragraph
2	Gaza rescuer	The Israeli military said Saturday it had launched "extensive strikes" as part of a fresh offensive in Gaza, after rescuers reported 100 people killed in the besieged Palestinian territory.
3	Gaza rescuer	The army said on Telegram it had begun the "initial stages" of the offensive, known as Operation Gideon's Chariots.
4	Gaza rescuer	The operation was part of "the expansion of the battle in the Gaza Strip, with the goal of achieving all the war's objectives, including the release of the abducted and the defeat of Hamas", it said in a post in Arabic.
5	Gaza rescuer	A separate statement in English said the army was "mobilizing troops to achieve operational control in areas of the Gaza Strip".
6	Gaza rescuer	Gaza's civil defence agency said Israeli strikes on Gaza had killed 100 people on Friday, while the army said its forces had "struck over 150 terror targets throughout the Gaza Strip" in 24 hours.
7	Gaza rescuer	Israel resumed its military offensive in Gaza on 18 March after a two-month truce in its war against Hamas, which was triggered by an attack by the Palestinian group in October 2023.
8	Gaza rescuer	The latest operation comes as Prime Minister Benjamin Netanyahu faces increasing pressure to lift a sweeping aid blockade on Gaza, as NGOs warn of critical shortages of food, clean water, fuel and medicines.
9	Gaza rescuer	The return to fighting since 18 March has drawn international condemnation, with the UN's rights chief on Friday denouncing the renewed attacks -- and what he described as an apparent push to permanently displace the population.
10	Gaza rescuer	"This latest barrage of bombs... and the denial of humanitarian assistance underline that there appears to be a push for a permanent demographic shift in Gaza that is in defiance of international law and is tantamount to ethnic cleansing," Volker Turk said in a statement.
11	Gaza rescuer	The main Israeli campaign group representing the families of hostages said that by extending the fighting, Netanyahu was missing a "historic opportunity" to get their loved ones out through diplomacy.
12	Gaza rescuer	Hamas on Friday demanded the United States press Israel to lift the aid blockade in return for a US-Israeli hostage released by the group.
13	Gaza rescuer	Edan Alexander, the last living hostage with US nationality, was freed last week after direct engagement with the Trump administration that left Israel sidelined.
14	Gaza rescuer	As part of the understanding with Washington regarding Alexander's release, senior Hamas official Taher al-Nunu said the group was "awaiting and expecting the US administration to exert further pressure" on Israel "to open the crossings and allow the immediate entry of hu
15	Gaza rescuer	Israel says its decision to cut off aid to Gaza was intended to force concessions from Hamas, which still holds dozens of Israeli hostages seized during the 7 October, 2023 attack that sparked the war.
16	Gaza rescuer	US President Donald Trump acknowledged on Friday that "a lot of people are starving" in the besieged Palestinian territory.
17	Gaza rescuer	"We're looking at Gaza. And we're going to get that taken care of," Trump told reporters in Abu Dhabi, on a regional tour that excluded key ally Israel.
18	Gaza rescuer	The Arab League is to meet in Baghdad on Saturday to discuss regional crises, with Gaza expected to be high on the agenda.
19	Gaza rescuer	United Nations chief Antonio Guterres will attend the summit, and Spanish Prime Minister Pedro Sanchez -- who has sharply criticised Israel's offensive in Gaza -- is expected to address it as a guest.
20	Gaza rescuer	The Hamas attack of 7 October, 2023 resulted in the deaths of 1,218 people on the Israeli side, mostly civilians, according to an AFP tally based on official figures.
21	Gaza rescuer	Of the 251 hostages taken during the attack, 57 remain in Gaza, including 34 the military says are dead.
22	Gaza rescuer	The health ministry in the Hamas-run territory said 2,985 people have been killed since Israel resumed strikes on 18 March, taking the war's overall toll to 53,119.
23	Israel launch	The Israeli military said Saturday it had launched "extensive strikes" in the Gaza Strip over the past day as part of the "initial stages" of a fresh offensive on the besieged Palestinian territory.
24	Israel launch	The strikes were part of "the expansion of the battle in the Gaza Strip, with the goal of achieving all the war's objectives, including the release of the abducted and the defeat of Hamas", Israel's army said in a statement in Arabic on Telegram.
25	Israel launch	Gaza's civil defence agency earlier said Israeli strikes on Gaza had killed 100 people on Friday.
26	Israel launch	The offensive, known as "Operation Gideon's Chariots", comes as Israel faces pressure to lift a sweeping aid blockade in return for a US-Israeli hostage released by Hamas.
27	Israel launch	Israel resumed its military offensive in Gaza on 18 March after a two-month truce in its war against Hamas, which was triggered by an attack by the Palestinian group in October 2023.
28	Israel launch	That assault resulted in the deaths of 1,218 people on the Israeli side, mostly civilians, according to an AFP tally based on official figures.
29	Israel launch	Of the 251 hostages taken during the attack, 57 remain in Gaza, including 34 the military says are dead.
30	Israel launch	The health ministry in the Hamas-run territory said 2,985 people have been killed since Israel resumed strikes on March 18, taking the war's overall toll to 53,119.
31	Israel launch	Israeli media reported on Friday that the military had stepped up its offensive in line with a plan approved by the government earlier this month, though there had not been any formal announcement of an expanded campaign.
32	Israel launch	The military said its forces had "struck over 150 terror targets throughout the Gaza Strip" in 24 hours.

Figure: 1

Displays the scraped_data.csv file, which is generated

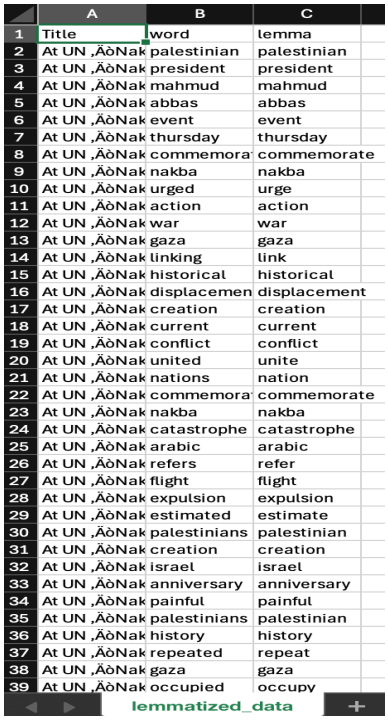
Text Cleaning and Tokenization

After scraping:

- Articles were grouped by title and paragraphs combined.
- Text was converted to lowercase.
- Tokenized using unnest_tokens().
- Common stopwords were removed.
- Numbers, punctuation, and short words were filtered.
- Lemmatization was applied using lemmatize_words() from textstem.

Clean tokens were saved in:

```
write.csv(clean_tokens, "lemmatized_data.csv")
```



	A	B	C
1	Title	word	lemma
2	At UN ,ÄôNak	palestinian	palestinian
3	At UN ,ÄôNak	president	president
4	At UN ,ÄôNak	mahmud	mahmud
5	At UN ,ÄôNak	abbas	abbas
6	At UN ,ÄôNak	event	event
7	At UN ,ÄôNak	thursday	thursday
8	At UN ,ÄôNak	commemora	commemorate
9	At UN ,ÄôNak	nakba	nakba
10	At UN ,ÄôNak	urged	urge
11	At UN ,ÄôNak	action	action
12	At UN ,ÄôNak	war	war
13	At UN ,ÄôNak	gaza	gaza
14	At UN ,ÄôNak	linking	link
15	At UN ,ÄôNak	historical	historical
16	At UN ,ÄôNak	displacemen	displacement
17	At UN ,ÄôNak	creation	creation
18	At UN ,ÄôNak	current	current
19	At UN ,ÄôNak	conflict	conflict
20	At UN ,ÄôNak	united	unite
21	At UN ,ÄôNak	nations	nation
22	At UN ,ÄôNak	commemora	commemorate
23	At UN ,ÄôNak	nakba	nakba
24	At UN ,ÄôNak	catastrophe	catastrophe
25	At UN ,ÄôNak	arabic	arabic
26	At UN ,ÄôNak	refers	refer
27	At UN ,ÄôNak	flight	flight
28	At UN ,ÄôNak	expulsion	expulsion
29	At UN ,ÄôNak	estimated	estimate
30	At UN ,ÄôNak	palestinians	palestinian
31	At UN ,ÄôNak	creation	creation
32	At UN ,ÄôNak	israel	israel
33	At UN ,ÄôNak	anniversary	anniversary
34	At UN ,ÄôNak	painful	painful
35	At UN ,ÄôNak	palestinians	palestinian
36	At UN ,ÄôNak	history	history
37	At UN ,ÄôNak	repeated	repeat
38	At UN ,ÄôNak	gaza	gaza
39	At UN ,ÄôNak	occupied	occupy

lemmatized_data

Figure: 2

Displays the lemmatized_data.csv file, which is generated

Using Document-Term Matrix (DTM) for Topic Modelling

The DTM is a structured format where each row represents a document (news article), and each column represents a unique word (term). The values indicate how often each word appears in a document.

We use the DTM because topic modeling (like LDA) works on numerical representations of text. DTM gives a quantitative view of text data that LDA can use to identify patterns and topics.

Code :

```
dtm_input <- clean_tokens %>%
  count(Title, lemma) %>%
  cast_dtm(document = Title, term = lemma, value = n)
```

Topic Modeling using LDA

LDA is a statistical model used to discover hidden topics in a collection of documents. It treats each document as a mix of topics and each topic as a mix of words.

We use LDA to automatically find topics in text data without labeling. It groups words that frequently appear together and assigns them to a topic.

Code:

```
lda_model <- LDA(dtm_input, k = 3, control = list(seed = 1234))
```

Output :

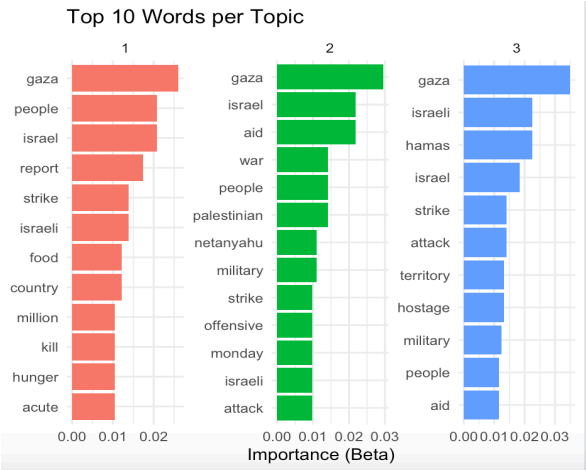


Figure: 3

Each topic is a theme the model found based on word usage patterns.

Top Terms per Topic
Each topic is defined by the top 10 most probable words (β values).

Model Evaluation

To decide how many topics (k) give the best result. Two common metrics:

- Perplexity: Lower is better. Shows how well the model predicts unseen data.
- Log Likelihood: Higher is better. Shows how probable the model's structure is.

To evaluate model performance:

Perplexity and Log-Likelihood were plotted against topic counts from k = 2 to k = 10.

```
ggplot(eval_df, aes(x = Topics, y = Perplexity)) + geom_line()
```

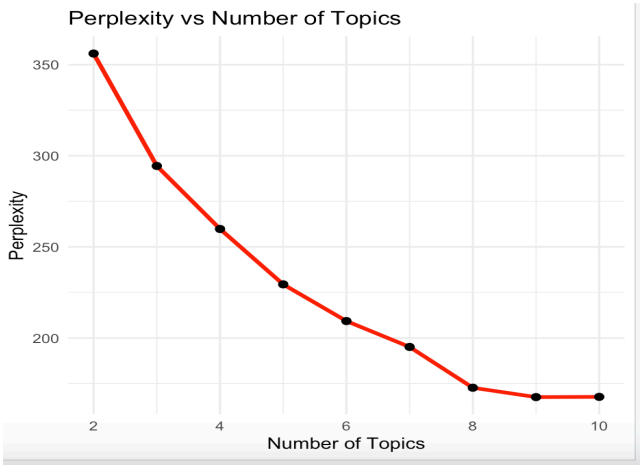


Figure: 4

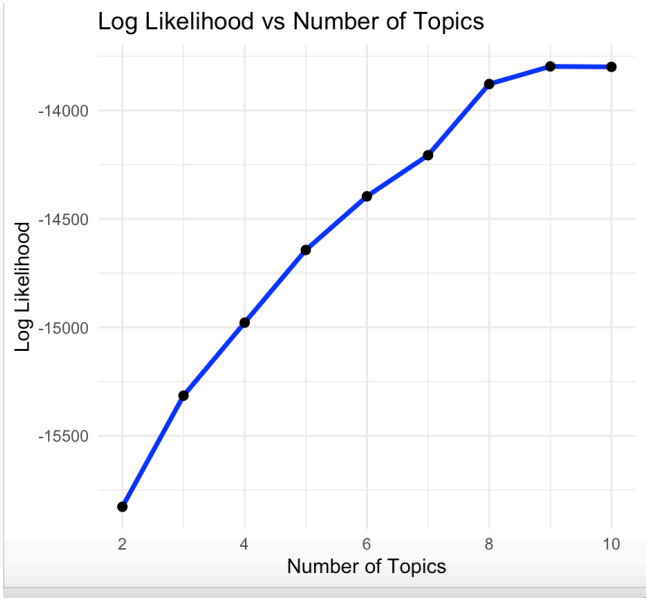


Figure: 5

These graphs help to identify the optimal number of topics (usually where perplexity plateaus).

Visualization

Using LDAvis, we visualized topic distances and word relevance interactively.

serVis(json_lda, open.browser = TRUE)

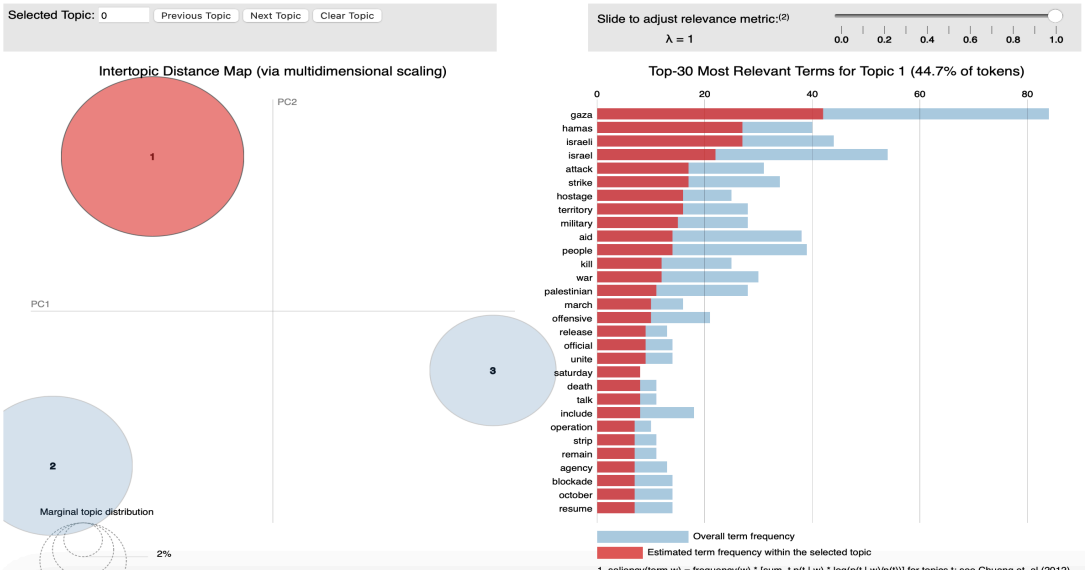


Figure: 6

Word Clouds

Generated one word cloud for each topic using wordcloud() to visually display high-probability terms:

wordcloud(words = topic_words$term, freq = topic_words$beta, ...)

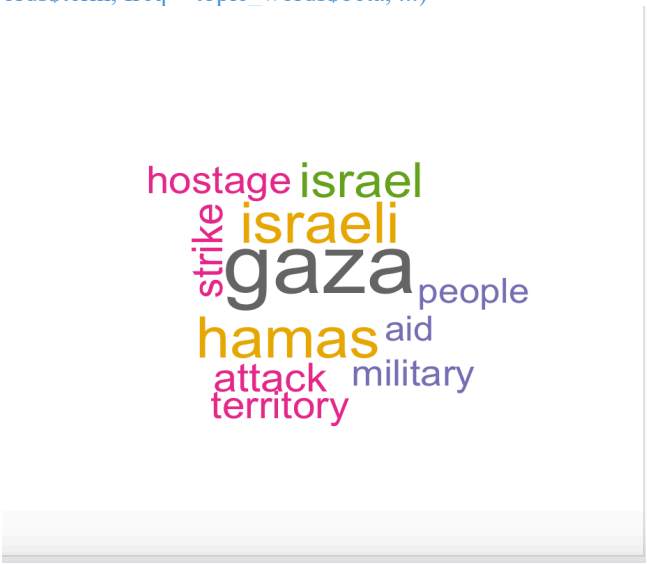


Figure: 7

Conclusion

We successfully scraped, cleaned, and analyzed online text data using R. By applying LDA, we identified key themes discussed in Middle East news articles. The project demonstrates how web scraping and topic modeling can be combined to derive insights from unstructured text. The process involved handling HTML data, preprocessing for text mining, and implementing statistical modeling for thematic analysis.The final models and plots provide a powerful summary of public discourse on current international issues.