**TOPIC-MODELING (Mehtab)**

**PRESENTATION**

**1. Visualizing Media Focus and Thematic Saturation (Bar Chart of Top 10 Topics)**

A graph with red and white bars

AI-generated content may be incorrect.

**Method Description**

The visualization draws on topic modeling, a natural language processing technique designed to detect latent themes in large text corpora. Using a pre-labeled CSV output of articles processed through a topic model (likely LDA), each article was assigned a topic number and accompanied by its top four keywords. These topic numbers represent recurring patterns of co-occurring words across documents. To prepare the data, articles labeled with topic “-1” were excluded, as these typically represent uncategorized noise or junk topics. Additionally, to enhance thematic coherence, topics whose top four keywords included three or more personal pronouns (e.g., “he,” “she,” “we”) were also filtered out. This step removed vague or conversational clusters unlikely to reveal coherent media themes. The cleaned data was then aggregated by topic to calculate frequency counts and percentages, ultimately selecting the ten most frequently occurring topics.

**Critical Reflection on the Method**  
Topic modeling, while not a perfect representation of meaning, is particularly well-suited to exploratory media analysis. It provides a macro-level view of thematic saturation—how often certain frames, regions, or actors appear. By combining it with keyword inspection and frequency counts, this method allows us to make meaningful inferences about which narratives dominate news coverage. One must be cautious, however: topics are probabilistic groupings, not definitive labels, and can reflect overlap or misclassification. Still, in this context, the method effectively highlights what themes persistently occupy media space, aligning with the project’s core interest in structured narrative repetition.

**Explanation of Visualization and Connection to Argument**

The horizontal bar chart ranks the ten most frequent topics in the news corpus by article count. Each bar is labeled with the top four keywords and the topic’s percentage share of total articles. The dominant topic—bank, west, israeli, Palestinian, alone accounts for over 12% of articles, suggesting a concentrated focus on the Israel-Palestine conflict. Other high-frequency topics include captives, hamas, release, hostages, hospital, patients, medical, hospitals, and iran, iranian, syria, us.

The design choice to use a horizontal bar chart serves clarity: longer topic labels remain readable, and frequencies are directly comparable. The red color emphasizes saturation and intensity of coverage. This visualization supports the project’s argument by making visible the hierarchy of attention in media reporting. It exposes how certain themes, especially those involving conflict, humanitarian crises, or geopolitical actors, dominate public discourse. This is not thematically neutral reporting; the chart visualizes structured repetition and narrative prioritization, revealing how some subjects are continuously centered while others remain peripheral or absent.

**2. Tracking Temporal Shifts in Media Narratives (Line Chart of Top 5 Topics)**

A graph of a line graph

AI-generated content may be incorrect.  
**Method Description**

This visualization focuses on temporal dynamics using the same topic-labeled corpus. First, the date of each article was converted into a datetime object and grouped by month. For the top five most frequent topics (determined from the previous bar chart), monthly article counts were calculated. To reduce noise and reveal underlying trends, a 3-month rolling average was applied to each topic’s count. This smoothing approach is widely used in time series analysis to highlight gradual trends while suppressing sharp fluctuations caused by anomalies or short-term bursts.

Critical Reflection on the Method

Tracking topic frequencies over time is essential to understand narrative emergence and decline. The rolling average is a well-suited technique for this purpose, it enhances legibility without distorting the data’s general shape. Still, limitations remain: topic labels do not evolve over time, so semantic drift or changes in vocabulary may be missed. Also, using only five topics is a design compromise - too many lines would obscure clarity, but fewer may leave out important shifts. Overall, the method effectively balances interpretability and insight.

**Explanation of the Visualization and Connection to Argument**

The line chart visualizes the evolving presence of the top five topics across the corpus timeline. The X-axis tracks publication month, while the Y-axis shows average article count (3-month rolling). Different colored lines represent distinct topics (e.g., captives, hamas, release, hostages or gaza, people, killed, younis).

The most striking insight is the spike in coverage for several topics around late 2023, aligning with the escalation of the Israel-Gaza conflict after October 7. For example, the gaza and hospital topics surge sharply, likely reflecting intensified humanitarian reporting. This temporal clustering supports the project’s thesis that media attention is event-driven and selective. Some themes lie dormant and then sharply rise in prominence, only to taper off again.

Design-wise, the use of color-coded lines allows topic differentiation, while the rolling average ensures the visualization captures narrative momentum rather than erratic reporting. The chart provides a temporal lens through which we can understand how specific events catalyze thematic saturation—a core pillar of the overall project argument.

**EXPLORATION**

**1. Distribution of Topics in Corpus (Excluding -1)**

A graph with a number

AI-generated content may be incorrect.

**Method and Critical Awareness**  
This bar chart was generated by counting how many articles were assigned to each topic number after running an unsupervised topic modeling process on the news corpus. Topic -1, which represents unclassified or noise-heavy articles, was excluded to clarify the distribution. Each article is tagged with the most probable topic based on its word composition.

**What the Visualization Shows**  
The bar chart shows a steeply uneven distribution: a handful of topics dominate, with Topic 0 appearing in over 350 documents, while many others occur in far fewer. This insight demonstrates that a small number of topics account for a large portion of the discourse, revealing thematic saturation. It also provides an empirical basis for filtering decisions made in later stages—excluding sparse topics and pronoun-heavy clusters helped sharpen the final analysis.

This chart supports the claim that media coverage is not evenly distributed across themes. Instead, a few topics consistently dominate, guiding public attention. Recognizing this skew early on helped focus the project on media’s tendency to prioritize and repeat certain narratives over others.

**2. Topic Trends Over Time (Monthly)**

**A graph of different colored lines

AI-generated content may be incorrect.**

**Method and Critical Awareness**

This line chart was produced by grouping the topic-labeled articles by month and plotting how many times each topic appeared. The time resolution was kept monthly to show short-term fluctuations, revealing how topics gain or lose attention in sync with real-world events. While this method is useful for identifying peaks in media focus, it can also produce noise, especially with less frequent topics that may clutter the plot.

What the Visualization Shows

Each colored line represents a topic. A sharp cluster of spikes is visible around late 2023, indicating a significant surge in coverage of a few dominant themes. The chart clearly visualizes temporal clustering in topic volume—certain narratives rise rapidly in response to triggering events (e.g., conflict escalation). However, the chart also reveals that many topics remain low and relatively flat throughout the period, underscoring their limited presence.

Connection to Project Argument

This visualization directly supports the argument that media attention is event-driven and thematically selective. The dramatic but uneven spikes in coverage confirm that certain topics absorb attention intensely for short bursts, shaping how the public experiences and remembers key events. This informed the decision to later use smoothing (rolling averages) for more legible trend analysis.

**Top Keywords per Topic**



To better understand the themes recurring across the news corpus, I produced a visualisation of the top four keywords associated with each topic. This was done by identifying the most frequent words assigned to each topic across the dataset. The result is a table that offers a quick thematic summary of the 79 topics generated from the news articles.

The table shows, for example, that topics like “captives, hamas, release, hostages” and “gaza, people, killed, younis” dominate certain clusters, suggesting a sustained focus on hostages and casualties. Others such as “protesters, palestine, police, protest” reflect civic responses to the conflict, while “unrwa, funding, agency, refugees” highlights humanitarian angles.

This visualisation ties directly into the main argument of the report: that media coverage tends to repeat and prioritize certain themes, shaping public discourse through consistent emphasis. By looking at which keywords consistently co-occur in topic groupings, we get a sense of how specific narratives—such as conflict, diplomacy, protest, and aid—are ranked and reinforced over time.

While keyword-based topic summaries can overlook nuance, they are useful for revealing broad patterns in how media constructs its narratives. In this case, they offer a strong foundation for tracing the **media’s thematic saturation** and support the overall claim that public attention is steered by recurring frames rather than neutral reporting.