**INTRODUCTION**

This project investigates types of themes, language and structures that are present in a large sample of Al Jazeera articles about war in Gaza, and questions how media makes sense of ongoing conflict longitudinally. Our group used four computational text analysis techniques - topic modeling, TF-IDF similarity, n-gram frequency and correlation with article length - to evaluate how media attention, language usage and editorial decisions changed in war time reports.

The dataset includes thousands of articles composed between 7 October 2023 and today, and reflects the fluid relationship between conflict, diplomacy, humanitarian catastrophe and geopolitical maneuvering. While other media studies have explored framing and bias with close reading, we employed what has been called “distant reading” to identify patterns that are otherwise invisible when scaled. Each technique pursued a different facet of news discourse: topics via topic modeling to expose the dominant and repeated themes, similarity via TF-IDF to cluster together sets of very similar articles, news trends via n-grams to track changing semantic frames and article length to assess the depth and editing weight associated with various topics.

Cumulatively, these discoveries support a central contention: media representations of war do not reflect a neutral recording of events but rather a patterned story marked by repetition, emphasis, and evolution in emphasis. Some issues are chronically prevalent, while others ebb and flow according to the political and humanitarian demands/requirements on the ground. The methodologies applied in this project enable us to see such patterning of language, attention, and thematic focus over time - as this is indeed how much news discourse constructs public understanding of conflict.

**ARTICLE LENGTH AND TOPIC MODELING ANALYSIS – BY RUBICA SHAH**

By merging two datasets—one with monthly average article lengths and another with topic proportions derived via Latent Dirichlet Allocation (LDA)—I investigate whether certain topics require more detailed coverage. My central argument is that some topics consistently correlate with longer articles, and that shifts in topic prominence reflect changing editorial priorities.

Method

I used two data set:

· length-year-month.csv: This file contains the average article length (in words) for each month in the al-Jazeera Gaza corpus.

· topic-model.csv: This dataset contains monthly topic proportions derived using **Latent Dirichlet Allocation (LDA)**.

Using **Python** and the **pandas** library, I performed the following data preparation and analysis. After converting the year and month columns into a unified datetime format (with day=1), I merged the datasets on this column. I then:

• Merged datasets on date

• Cleaned using stopwords

• Extracted dominant topics

•Found the top topic for each month using .idxmax()

• Created visualisations using plotly (line chart, scatter plot, and bar chart)

**N-GRAM – BY RUKHSHAN REHMAT**

For this project I used the ngram data frame and used 2-gram-year-month.csv dataset. As the 2-gram year month csv had all the bigrams through tokenization used in AL Jazeera news articles, aggregated by year and month. I took 2 themes so that I could really categorize the data aggression-related (e.g., air strike, civilian casualties) and sympathy-related (e.g., humanitarian aid, cease fire). I removed stop words using NLTK, group-ed monthly counts with pandas, and calculated relative frequencies to allow fair comparison across months with different article volumes. Finally, I visualized the results using Plotly Express, creating a faceted stacked bar chart that showed the rise and fall of each category over time. This method allowed for clear temporal comparison and made it possible to track how media language shifted between aggression and sympathy across the conflict timeline.

**TFIDF SIMILARITY ANALYSIS – BY SAHAR MUBEEN**

I used TF-IDF and cosine similarity to identify the articles in the Gaza war corpus that are thematically similar in terms of their vocabulary patterns. The most similar articles in pairs show how specific subjects, such as hospital bombing and refugee crises, are discussed frequently and with similar language in several articles. This indicates central themes in media coverage.

Using TF-IDF and cosine similarity of this component, I investigated lexical overlap between the articles of the Al Jazeera Gaza corpus. I developed both the exploration and presentation scripts for TF-IDF.

The dataset “tfidf-over-0.3.csv” had precomputed cosine similarity scores between article pairs based on their TF-IDF vector representations. It is a widely used method in computational text analysis which down-weights phrases that appear frequently. It also prioritizes words that are rare to specific documents. This made it an appropriate method for finding meaningful thematic overlaps within our corpus.

In the exploration step, I loaded the dataset using pandas, verified data integrity, then examined the distribution of cosine similarity scores. I separated article pairs with scores higher than 0.7 for further analysis. Assuming that the articles are quite similar in terms of subject matter.

In the presentation part, to create a bar chart I used plotly express. The chart shows the top 10 similar article pairs. This visualization made it simple to find recurring themes, E.g hospital bombing and ceasefire agreements. These themes were often covered in more the one article, at times appearing in almost identical language.

Seeing these clusters represents how often certain vocabularies were repeated, whether intentionally or not, in the reporting of the Gaza conflict.

TF-IDF, even with its limitations, proved to be a powerful tool for this kind of large-scale exploration. Its speed and simplicity allow to quickly find out which articles might be talking about the same events. It helped me decide where to look more closely with a critical, human lens. For this project, TF-IDF wasn’t about replacing close reading; it was about knowing where to start.

**TOPIC MODELING – BY MEHTAB ALI KHAN**

The goal of this part of the project was to identify and illustrate the recurring themes in a sizable collection of articles, with an emphasis on how the media shapes and ranks subjects over time. I aimed to illustrate patterns of media attention and thematic saturation, that is, what subjects are regularly emphasized and how that focus changes in response to events, by examining topic modeling results and creating visual representations of topic distribution and temporal trends.

A topic-labeled CSV file from a corpus of news articles served as the dataset. To ensure data clarity, I filtered out all articles labeled with topic -1, which usually marks irrelevant or uncategorized documents. I also removed topics that were strongly skewed by pronouns.

The cleaned data was utilized to determine the top 10 most common topics, ordered by article frequency. These were represented in a horizontal bar chart to display which themes prevailed over the media corpus. To study narrative development, I also concentrated on the top 5 topics and graphed their article frequency by time via a 3-month rolling average. It smoothed out, enabling me to highlight long-term tendencies above short-term fluctuations and making changes in media attention easier to depict.

The central argument my findings make is that media coverage is not thematically neutral; it has recurring patterns of attention, saturation, and neglect. Some themes keep on occupying attention, framing public debate through iteration and prioritization. By monitoring these trends over time, my work proposes that media narratives are patterned and structured.

**VISUALISATIONS AND DATASETS**

**ARTICLE LENGTH AND TOPIC MODELING ANALYSIS – BY RUBICA SHAH**

*Critical Reflection*

The method helped reveal broad trends but had limitations. Many topic columns had missing values, and a generic (Topic) label appeared as the dominant one in 4,185 months, suggesting a possible coding or extraction error. This reduced the diversity of ideas.

While LDA is useful for large corpora, its topics are abstract without keyword examination. Also, merging data by month flattens article-level variation, possibly hiding finer-grained shifts in reporting.

*Visualisations and Interpretation*

**1. Dominant Topic Counts – Bar Chart**

A graph with a bar and text

AI-generated content may be incorrect.

The chart shows how many months each topic was dominant, but it shows a key issue: a generic label (topic) dominates the data, even though it’s not an actual topic. This likely happened because of a mistake in how the dominant topics were extracted, such as using idxmax() before the topic labels were assigned correctly. The complete absence of labels like (topic\_3) raises further doubts about the accuracy of the topic modeling. This supports the argument that without careful data cleaning and validation, the insights we draw, like which topics are most important, can be misleading or incorrect.

.

**Article length and topic proportions over time**

A graph with a line and a line

AI-generated content may be incorrect.

The line chart titled *"Article Length and Topic Proportions Over Time"* reveals a clear pattern in journalistic trends from 2016 to 2025. The average article length slowly increased until around 2020, suggesting a period of more in-depth reporting. However, this trend reverses post-2020, where a noticeable decline occurs. Interestingly, this shift coincides with a visible emergence and fluctuation of topic\_1, which only starts appearing significantly after 2020. In contrast, topic\_2(red color) maintains a relatively stable presence across the entire timeline. This relationship suggests that changes in dominant topics may be influencing the depth of news coverage, supporting the argument that topic shifts potentially driven by political or social change directly impact how thoroughly issues are explored in media narratives.

**Average Article Length vs Topic**

A graph with blue dots

AI-generated content may be incorrect.

The scatter plot titled *"Average Article Length vs Topic Proportion"* shows that article length remains relatively stable across varying topic intensities. Despite fluctuations in topic proportions, some reaching above 80, most articles consistently fall within the 650 to 700-word range. This lack of a strong linear relationship suggests that while topic prominence may shift over time, it does not significantly influence the length of articles. Such consistency in word count points to standardized editorial practices that prioritize readability over content-specific depth, strengthening the argument that institutional factors often override content variation in shaping article structure.

*Conclusion*

Combining topic modeling with article length reveals that editorial depth may shift with topic prominence, but only subtly. The presence of data issues like the "Topic" label error and missing values, limits interpretation. Nevertheless, the analysis supports the idea that editorial focus is dynamic, and certain topics may invite more detailed coverage over time.

**N-GRAM – BY RUKHSHAN REHMAT**

For my part in the project, I worked with n-gram data to study how language changes over time in media coverage of Gaza and Israel. I started by exploring the datasets for 1-grams, 2-grams, and 3-grams, each of which contained the number of times different phrases appeared in news articles, breaking down by month and year. My goal was to identify how the language used in reporting shifted during the war period. At first, I tried analyzing all three n-gram types together, if this would give me a more complete picture. However, I soon realized that combining them made it hard to find a consistent theme or message, since they produced a lot of noise and randomness.

A screenshot of a computer screen

AI-generated content may be incorrect.

1. Even when I focused just on **2-gram phrases**, the top 10 or even top 50 phrases by frequency didn’t reveal anything particularly useful. Many of them were generic (e.g., “west bank” , “prime minister”, “south Africa”, “Biden administration”) or out of context. At this point, I decided to switch strategies. Instead of relying on purely quantitative ranking, I went through the data manually and identified phrases that carried meaning. I noticed that some phrases like “air strike”, “civilian casualties”, “rocket fire”, and “military attack” appeared frequently, and clearly reflected aggression or violence. On the other hand, I also found bigrams like “humanitarian aid”, “cease fire”, “medical help”, and “international support” that conveyed sympathy and relief.

A blue and white graph

AI-generated content may be incorrect.

That inspired my final approach: to create two semantic categories — one for aggression-related phrases and another for sympathy-related phrases — and visualize how their usage changed over time. I used the 2-gram-year-month.csv file and filtered the data to include only those phrases that fit either category. I then used **pandas** to group and sum the counts per month, and calculated **relative frequencies** (i.e., each phrase’s proportion out of all phrases used in that month). This was important because some months had more articles than others, and raw counts alone could be misleading.

A screenshot of a graph

AI-generated content may be incorrect.

I decided to use Plotly Express and make a faceted stacked bar chart for this stage. In the upper chart, phrases connected to feelings of sympathy were shown and in the lower chart, phrases connected to feelings of aggression were displayed, every bar covering a month from 2021 to 2024. It allowed me to more easily spot the trends in every kind of language across time. From mid-2021 through early 2023, most of the popular phrases on the chart mentioned the war in aggression-related ways, especially those connected to “civilian casualties” and “air strike”. At the same time, violence has increased, which is typical at these stages. Nevertheless, by late 2023 and early 2024, using sympathy-related phrases became more common. News media mentioned “humanitarian aid” and “cease fire” more often, hinting that they were paying more attention to helping people. Thinking carefully, I observed that the language used in the media in a war zone undergoes changes. First, the focus is on reporting what was destroyed and attacked, while later, the focus is on individuals affected and the global community’s actions. Apart from demonstrating the types of languages used, my project displayed the periods when various languages became common. For instance, language expressing conflict became very strong during the first weeks, while feelings of sympathy became more frequent afterwards, usually triggered by violence. This shows that sympathy interest in media only appears when aggression goes too far. This might show how news groups react to people’s feelings or how pressures from other nations form after a bad event occurs.

This method has its limitations. Working with 2-grams doesn’t provide the full sentence context and not even the semantics or sentiment so it’s hard to be 100% sure how the words were used. Also, some phrases were — for example, “strike” might refer to military action or a worker protest and removing stop words were also changing the semantics and adding them makes them meaningless. Despite these challenges manually categorizing the bigrams gave me a clearer view than relying on frequency alone makes no sense when analysis such a large corpus.

**TFIDF SIMILARITY ANALYSIS – BY SAHAR MUBEEN**

**Top 10 Most Similar Articles Based on TF-IDF Cosine Similarity**

A screenshot of a computer

AI-generated content may be incorrect.

**High Similarity Scores**:

The bar chart titled "Top 10 Most Similar Articles Based on TF-IDF Cosine Similarity" presents a clear visualization of the most thematically aligned article pairs within the Gaza war corpus. Each article pair showed a high cosine similarity score, approaching 1.0. This indicates that these articles share nearly the same vocabulary and thematic focus.

This high degree of similarity underscores the presence of consistent narrative patterns across multiple media outlets. Particularly, dates such as 2023-12-01, 2024-03-22, and 2024-03-29 appear more than once among the top pairs. This suggests that news coverage around these times was heavily centered on recurring humanitarian issues, including civilian casualties and infrastructure destruction.

Articles such as:

* 2023-12-01\_1993.txt
* 2024-03-29\_225.txt
* 2024-03-17\_338.txt

show up in multiple high-similarity pairings, pointing to possible content replication or syndication across different sources. This visualization strengthens the central argument of the report that media coverage of the Gaza war is characterized by a concentrated use of recurring language and themes.

**Media Framing and Thematic Clustering**:   
 The graph highlights a limited thematic range dominating Gaza war coverage, likely involving repeated mentions of

* **Hospital bombings**
* **Civilian casualties**
* **Displacement/refugees**

The analysis offers a deeper understanding of how media narratives are shaped and shared during times of crisis and how certain themes dominate public conversation.

**Conclusion**

The graph underscores how media narratives around the Gaza war tend to recycle key humanitarian themes using almost identical language and structure. This repetition may help emphasize urgency but can also signal limited narrative diversity, raising questions about how thoroughly and independently the conflict is being covered.

**TOPIC MODELING – BY MEHTAB ALI KHAN**

**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, 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 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**

A screenshot of a computer

AI-generated content may be incorrect.

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.

**CONCLUSION**

Our multi-method analysis of the Al Jazeera Gaza corpus validates that media war reporting is selective and patterned. This general statement is implied by each of the analytic parts. Among the topic modelling work, it demonstrated which themes receive attentions too much and less to much and less, and how these attentions fluctuate with events occurring in real world. TF-IDF similarity scores showed repetitive reporting on main crises such as hospital attacks, or ceasefire talks suggesting editorial prioritization of narratives. The n-gram frequency analysis mapped a semantic transition from the violent vocabulary of aggression-related to discourse of emergencies, illustrating how the media language accommodates a changing conflict. Finally, the analysis of article length has shown that types of topics related to complex humanitarian or geopolitical issues tend to receive longer and more detailed coverage, editorial decisions on the depth that some topics deserve.

By combining these approaches, our report can move beyond single instances of representation and toward a fuller picture: the contours of news coverage are made up of discernible patterns of attention, framing, and repetition. Computational techniques enable us to follow these patterns through history and reveal them as part of a broader web of connections between war and media representations. This work, ultimately, encourages a more skeptical-eyed reading of news, for reporting is as much about narrative making as it is about information and reportage, a collection of editorial decisions and institutional routines.