Project report

for

Digital Humanities

Course instructors

Mathew Barber

Peter Verkinderen

By :

Faizan Amir

Zia Ullah

Arsalan Danish

Kamil Ahmed

Introduction:

This project examines Al Jazeera's coverage of Gaza through computational text analysis, aiming to understand the patterns in how media narratives evolve over time and respond to key events. By combining four analytical methods article length analysis, n-gram frequency tracking, TF-IDF cosine similarity analysis, and topic modeling we explore both structural and thematic trends in the reporting.

Article length analysis reveals shifts in editorial priorities, such as changes in depth and frequency of coverage. N-gram frequency identifies the most prominent terms and phrases, shedding light on recurring themes. TF-IDF similarity measures how closely articles align in content, highlighting periods of repetitive or distinct reporting. Finally, topic modeling categorizes articles into broader discourses, such as conflict, diplomacy, and humanitarian issues, to assess narrative balance.

Through this multi-method approach, we provide a data-driven perspective on how Al Jazeera frames the Gaza conflict, offering insights into the interplay between media representation and real-world events. The project also establishes a clear, reproducible methodology for similar studies in digital humanities and media analysis

Documentation:

This project investigated Al Jazeera's coverage of Gaza through four methods: article length analysis, n-gram frequency tracking, TF-IDF similarity analysis, and topic modeling. Each team member contributed a unique component to the overall analysis, with the shared goal of identifying how the media narrative changed across years and in response to significant events. This document outlines the steps followed by each team member and provides a reproducible framework for anyone wishing to conduct similar research using updated or new data.

Faizan Amir led the article length analysis. His objective was to examine whether the word counts of articles changed over the years, potentially reflecting shifts in editorial strategy. He began by analyzing a dataset that included the average and total word counts of articles per year. To assess whether changes in total word count were due to article length or quantity, Faizan developed a script that counted the number of article files per year. Each article was named using the format YYYY-MM-DD-XXXX.txt, allowing him to extract the year and create a bar chart of article counts over time. The data revealed that 2021 had the fewest articles (excluding an outlier year, 2017), yet some of the longest ones. Faizan further analyzed the shortest and longest articles each year, finding that while the shortest articles remained relatively stable in length, the longest ones varied significantly. This explained the spike in average article length in 2021. His visualizations included a bar chart of articles per year and a line chart showing trends in article length extremes. This analysis provided insights into whether reporting leaned more toward in-depth features or concise summaries over time. To replicate this work, one must preserve the filename format and use word counts (not characters) for consistent comparison.

Kamil Ahmad focused on n-gram frequency analysis to uncover dominant words and phrases in the articles. He analyzed datasets of 1-grams (individual words), 2-grams (word pairs), and 3-grams (three-word phrases), verifying the formatting of year columns and ensuring the data was clean and consistent. Initially, his results were overwhelmed by common stop words like "the" and "and", so he implemented a standard stop-word filter to remove these and emphasize more substantive terms. Post-filtering, significant terms such as "Gaza", "Israel", and "Palestinian" emerged, providing a clearer picture of the coverage's thematic focus. Kamil created two main visualizations: a line chart showing frequency trends of top 1-grams across years, and a bar chart comparing overall usage of the top 10 terms. He also examined 2-gram and 3-gram patterns, identifying frequent phrases like "humanitarian aid" that shed light on how topics were framed. These visualizations revealed both stable and event-driven spikes in language use. To reproduce his analysis, users should ensure datasets include correctly formatted year columns, apply a robust stop-word list, and normalize term frequencies when comparing across years with uneven article counts.

Zia Ullah conducted the TF-IDF similarity analysis to evaluate how similar articles were to one another in terms of content. To ensure meaningful analysis, Zia began by excluding articles shorter than 200 words. He applied TF-IDF vectorization to convert each article into a numerical representation that prioritized rare but informative words over frequent ones. Using cosine similarity, he computed the pairwise similarity between articles, generating scores between 0 (completely dissimilar) and 1 (identical). Most similarity scores fell between 0.3 and 0.5, as shown in a histogram, while scores above 0.6 were uncommon. A heatmap focused on the 0.4–0.9 range revealed that high-similarity articles tended to cluster around specific time periods, indicating repetitive coverage during major events. This analysis showed how editorial practices vary, with certain events prompting more uniform reporting. To replicate this method, future researchers should apply the same length filter, use TF-IDF vectorization, calculate cosine similarity, and produce visualizations that focus on mid-to-high similarity scores for interpretability.

Arsalan Danish was responsible for topic modeling using Latent Dirichlet Allocation (LDA). The preprocessed dataset contained articles that had been assigned to machine-generated topics, each labeled with a set of associated keywords. Initially, Arsalan created a bar chart showing the most frequent topics, but these numerical topic labels lacked interpretability. To address this, he manually reviewed the keywords and grouped them into three thematic categories: "Safety and Conflict" (e.g., "Hamas," "military," "attack"), "Politics and Diplomacy" (e.g., "Biden," "UN," "peace talks"), and "Aid and Civilian Toll" (e.g., "hospital," "refugees," "children"). He then created a stacked bar chart visualizing monthly topic distributions, allowing comparisons of how thematic focus shifted over time. The analysis revealed that conflict-oriented topics dominated coverage, while humanitarian themes appeared more sporadically. To reproduce this work, one should begin with LDA output files containing topic assignments and keywords, manually group

topics into meaningful categories, remove generic terms (e.g., pronouns), and normalize data when article volume varies across months.

To maintain consistency across all analyses, team members ensured that datasets were aligned by time and maintained a clear folder structure. Raw articles, n-gram datasets, LDA outputs, classified topics, and visualizations were kept in separate directories for clarity. This documentation provides a help for reproducing the project with new data and offers a varigated understanding of how news narratives surrounding Gaza were constructed, repeated, and emphasized by Al Jazeera over time.

Individual Reports:

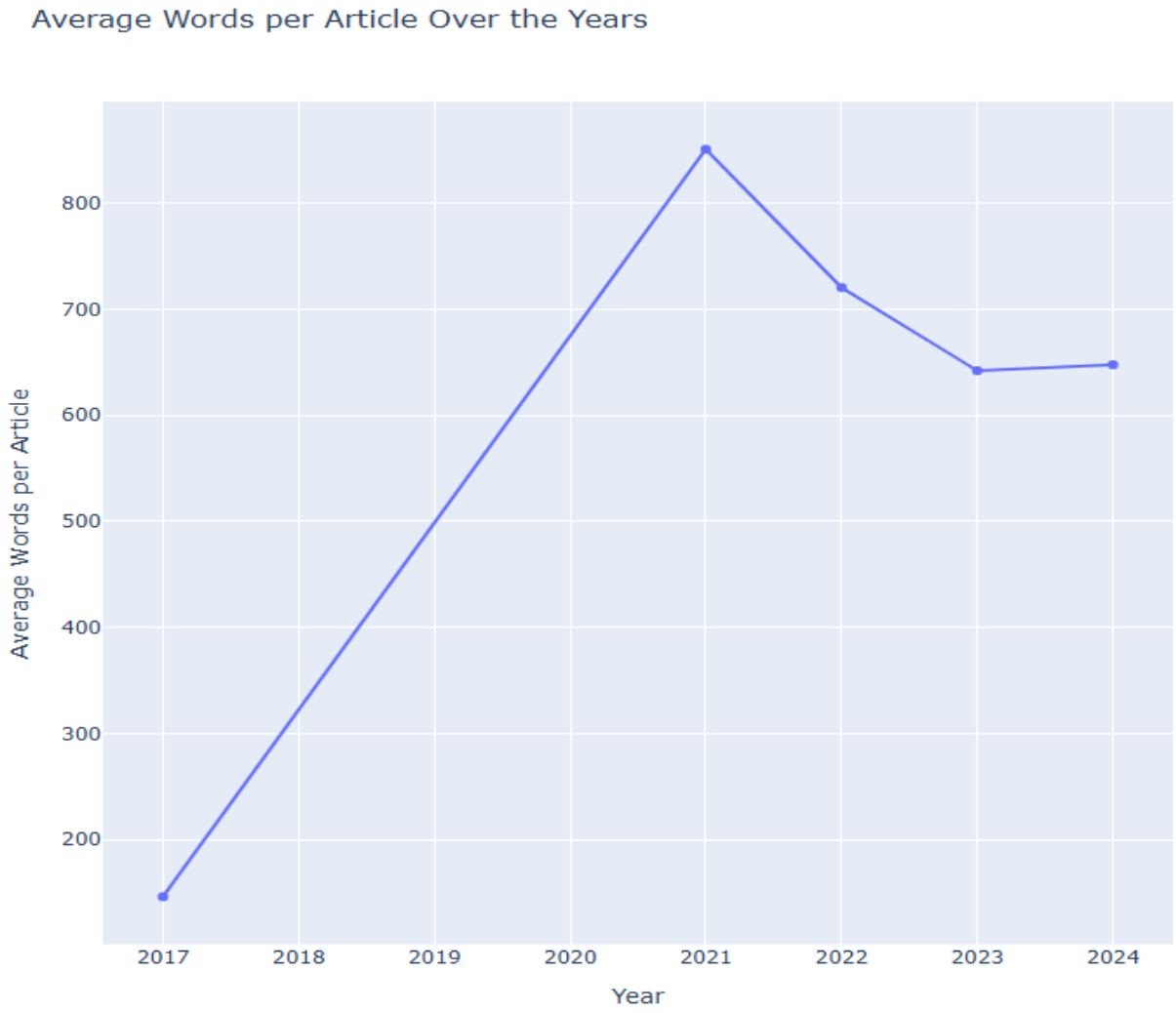
FAIZAN AMIR

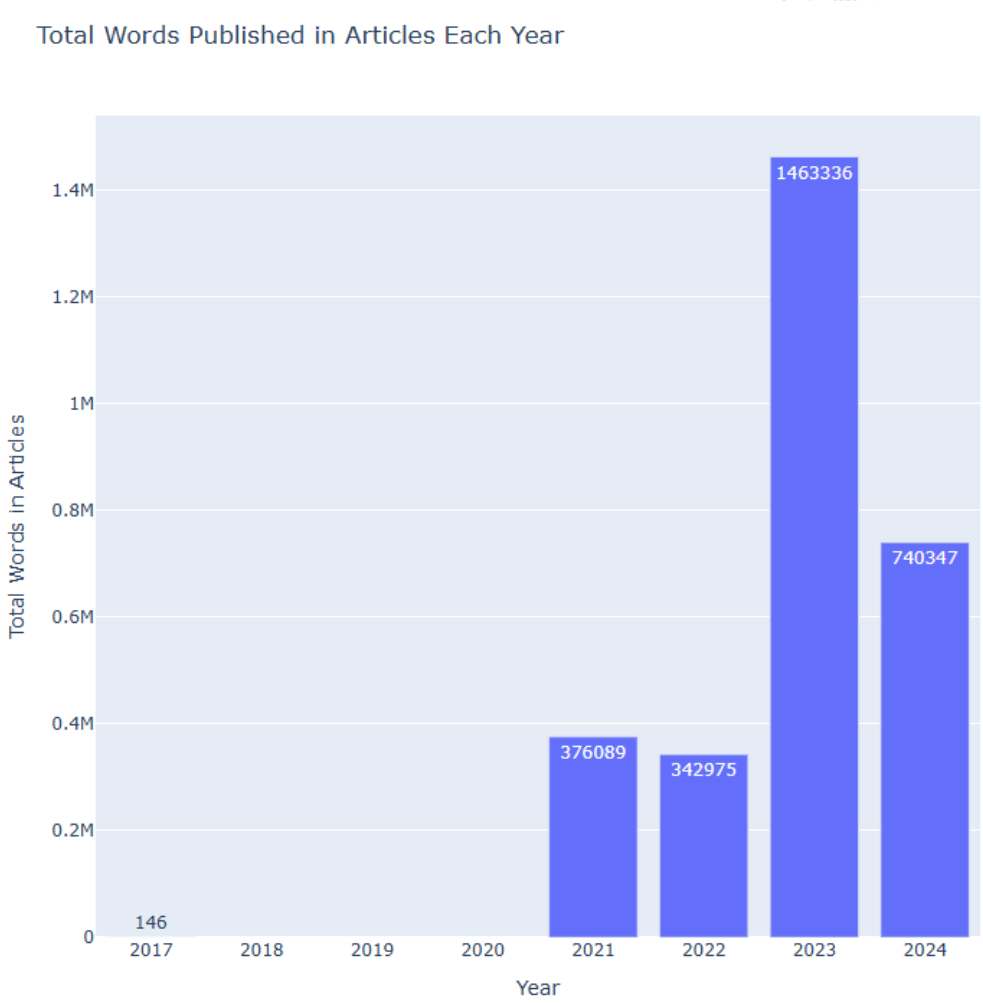
Article length

Article length refers to how long a text is, usually measured by the number of words, characters, or sentences it contains. In digital humanities, analyzing article length helps us compare texts, study writing patterns, or visualize trends in large datasets. For example, shorter articles may summarize key ideas, while longer ones might include detailed analysis. By measuring and visualizing article lengths, we can better understand differences in writing styles, genres, or publishing trends.

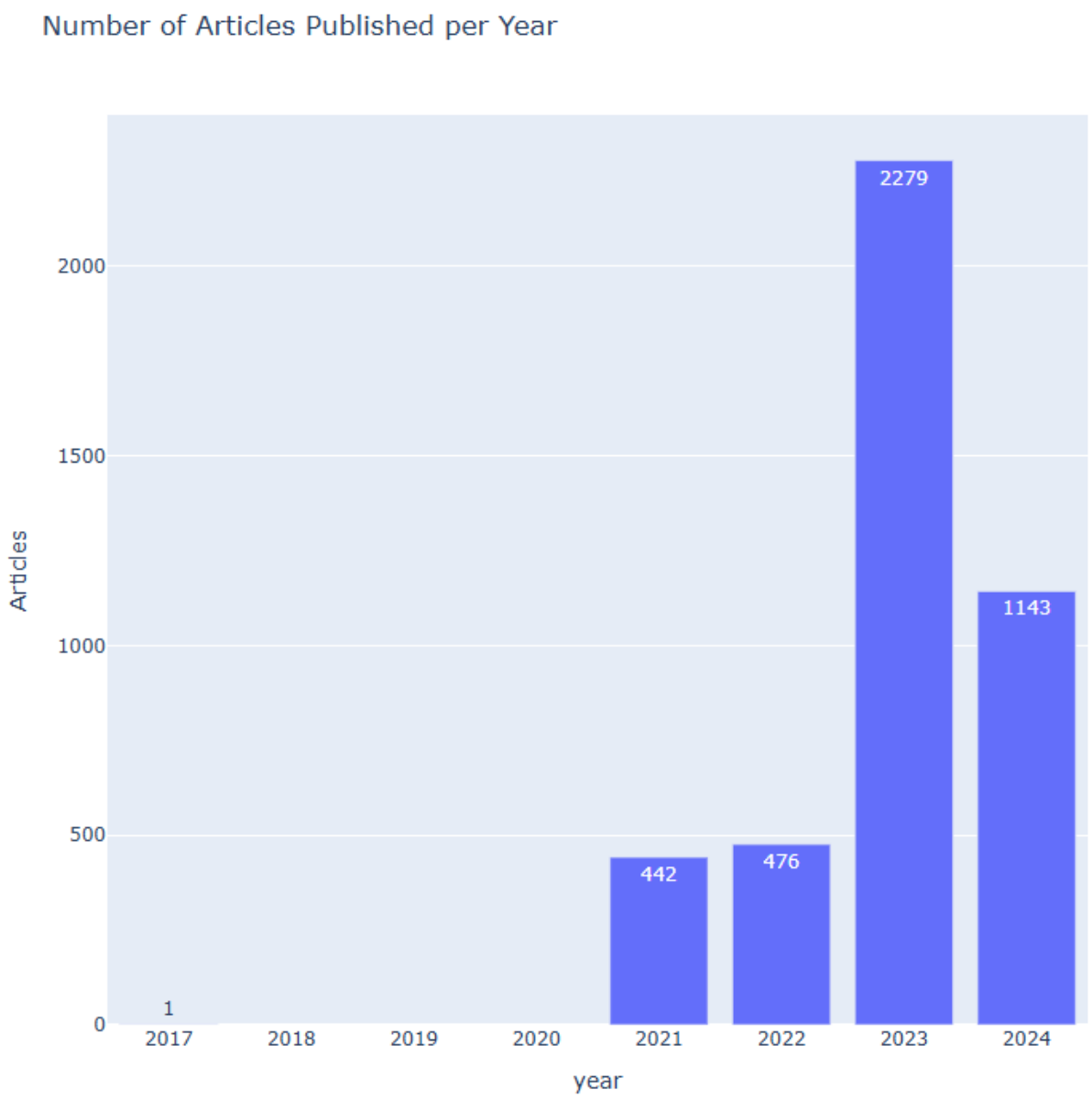
While exploring the article length dataset, I focused on the length-year DataFrame, which included the total word count and average word count per article for each year. It is important to know that average and total word counts are commonly used in text analysis as basic structural indicators. They don't reflect content quality or meaning, they serve as useful proxies for editorial style across time. I created two separate visualizations: one showing the total number of words per year and another displaying the average word count per article per year. I observed that the average words per article peaked in 2021, whereas the total number of words was highest

in 2023. This made me question what might explain these trends.



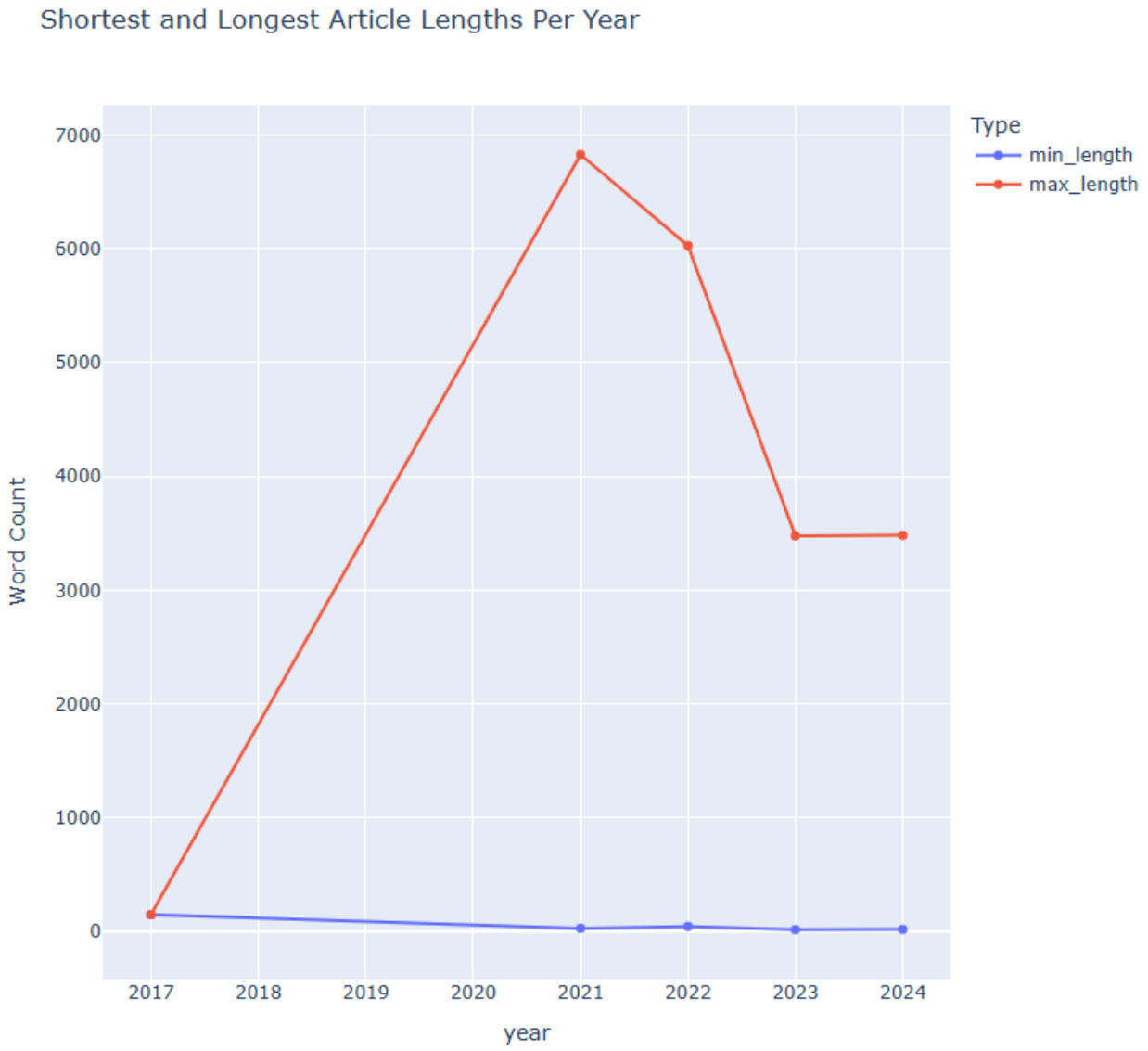


To investigate further, I assumed whether each year had the same number of articles, which would affect both total and average values. However, the original DataFrame did not include the number of articles per year. To address this, I turned to the article folder containing all the .txt files. Since the files were named in the format YYYY-MM-DD-XXXX.txt, I extracted the year from each filename and created a new DataFrame to count the number of articles per year. I then visualized this in a bar chart, which clearly showed that the number of articles varied significantly year by year. This disproved the assumption that article counts were constant across years. This method works if all filenames are properly named and formatted. Later, it would be good to double-check that all files follow the correct naming rule to make sure no mistakes or missing files affect the results.



Given that 2021 had the highest average article length, I hypothesized that its articles were fewer but longer, while 2023 had more articles, possibly shorter in length. To test this, I wanted to compare the minimum and maximum article lengths across years. Since this data wasn't available in the original dataset, I wrote a script to read each article file, count the words, and group these word counts by year. For each year, I identified the shortest and longest article and stored the results in a summary DataFrame. I then created a line chart to visualize how the minimum and maximum article lengths changed over time. Word count is an easy way to measure article length, but it has some downsides. For example, it might include things like headings, formatting, or metadata, which aren’t part of the main text. It also doesn’t show how

complex or varied the content is. Still, for structural trends, it remains a valid approach.



From the visualization, I observed that the shortest articles each year were approximately similar in length, which showed that there is a consistent lower bound on how short articles tend to be. However, there was noticeable variation in the longest articles, particularly with 2021 having one of the longest maximum article lengths compared to other years. This supports the assumption that 2021 had fewer but much longer articles, which would explain the spike in average article length for that year. On the other hand, while 2023 had the highest total word count, its maximum article length was not as high, reinforcing the idea that 2023's average was lower because it had more but relatively shorter articles. This pattern clarifies the contrasting trends observed in the total and average word count visualizations. This suggests that extreme values (e.g., exceptionally long articles) can significantly increase yearly averages, even if the number of articles is lower. The presence of such outliers plays a key role in shaping summary statistics like the mean. All these arguments show that the articles in the corpus which were written in

2021 are longer in length but fewer in number than other years which can be due to the content or the trend of writing in that particular year which needs to be further analyzed linking it with other text analysis techniques like TFIDF and topic modelling.

This process helped me explore patterns in article length more meaningfully and understand how both the quantity and size of articles contributed to yearly trends in total and average word count. Overall, this method is appropriate for identifying structural variations across time but does not support more interpretive claims about the content itself. Future work might pair this with more advanced text analysis technique such as topic modeling for a content-based understanding.

N-Gram (Kamil Ahmad)

For my contribution to the project, I was responsible for analyzing the n-gram dataset, which contains frequency counts of words (1-grams, 2-gram and 3-gram) used in Al Jazeera articles related to Gaza, organized by year.

The objective of my analysis was to identify the most frequently used content words in the corpus over time, visualize their trends.

Although I explored all the data sets (1-gram, 2-gram and 3-gram), particularly gram-year file in each of the three files, but for the final presentation script I choose 1-gram-year data set.

I visualized all the yearly data sets of 1-gram, 2-gram and 3-gram and tried bar chart and line chart. And, for the 1-gram-year initially I visualized without considering the stop words and it turned out to be meaningless because it was counting frequencies like (at, the , at) etc. but for the final presentation and other exploration I removed the stop words.

I began by loading and inspecting the dataset using pandas, checking for data integrity, and confirming that the year column was properly formatted as an integer. To ensure meaningful results, I manually removed common stop words and non-informative terms which I took from a website (https://gist.github.com/sebleier/554280) before ranking the 1-grams by their overall frequency across all years.

After filtering, I selected the top 10 most frequent and meaningful 1-grams for further exploration. I produced two visualizations:

- A line chart showing the yearly frequency trends of these top words across the full-time span, and
- A bar chart displaying the total overall frequency of the top 10 terms.

These visualizations reveal how certain terms consistently dominate the reporting on Gaza, and how their usage intensity varies during key events. The line chart is particularly useful for understanding temporal spikes, while the bar chart provides a snapshot of overall prominence.

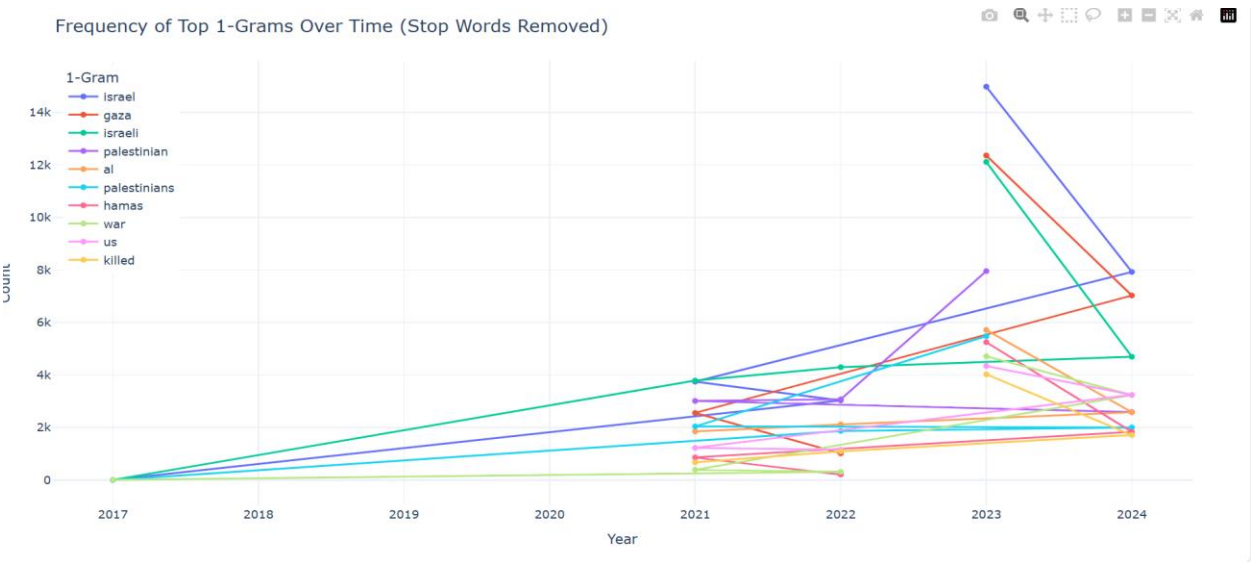
The final outputs were exported as HTML files.

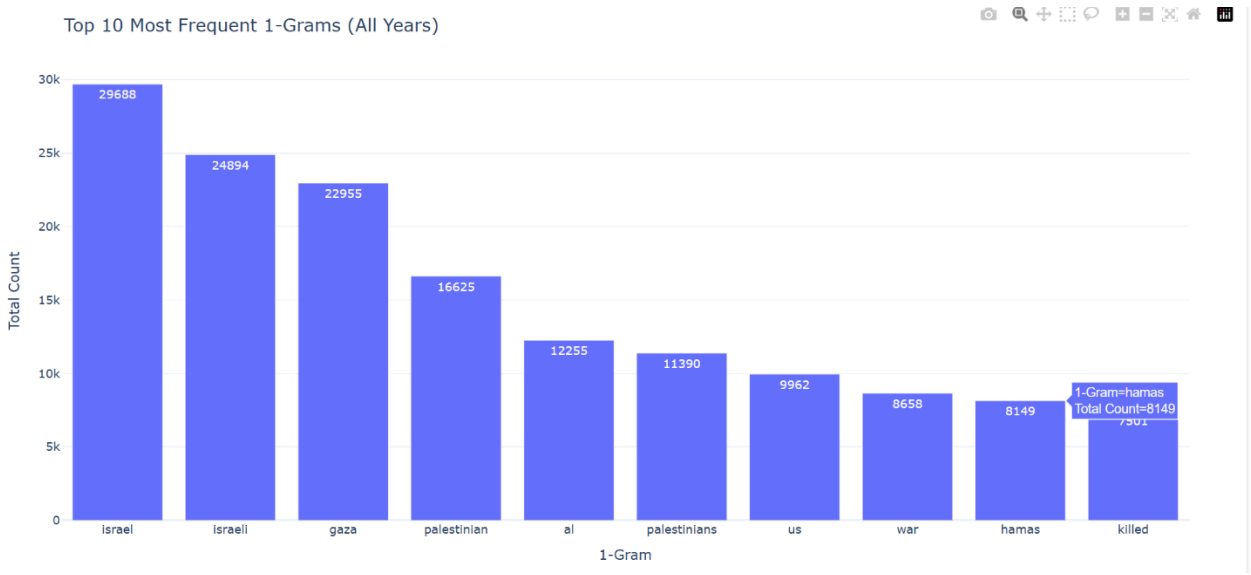
Methodology:

In this project, I concentrated on the n-gram datasets-1-gram-year, 2-gram-year, and 3-gram-year files. These are datasets showing how often single words (1-grams), pairs of words (2-grams), and three-word phrases (3-grams) occurred in Al Jazeera's Gaza related articles over the years. I started with the 1-gram-year data and created a visualization that included all words. However, I quickly realized that most of the top results were stop words like "the," "and," and "of," which don't provide meaningful insight. So, I removed these stop words using a standard list to focus on more important words. I then tried creating a bar chart with the cleaned data, but it didn't look very clear or informative for showing changes over time.

Next, I explored the 2-gram and 3-gram yearly datasets. These helped highlight common phrases rather than just individual words, and I visualized them using bar charts to show the most frequently used expressions across the entire time span. This gave a better sense of recurring themes in the media coverage. Although I worked with all three datasets, I chose the 1-gram-year data for my final contribution to the presentation. This time, instead of a bar chart, I created a line chart showing how the frequency of the top 1-grams changed year by year. This chart made it easier to spot patterns and spikes in word usage during major events. My work was to spot useful language shifts over time and aid in showing how news talk about Gaza has changed.

While seeing the top used words in Al Jazeera write-ups on Gaza, we found that terms like "Israel," "Israeli," "Gaza," and "Palestinian" came up the most. This shows that the story mostly centers on the key players in the clash. These words were used again and again over the years, and they increased during times of fighting or big events. Clearly, the evolution of the usage of these words during usage over time can be seen through integrated line and bar graphs. Thus, we have a very concrete understanding of how media have used the words, as well as how the media shape discourse around the Gaza conflict and its focusing.





Using n-gram frequency is a quick and easy way to have a look at a very large number of articles to see which words crop up most of the time. It helps us get a general idea of common themes and is simple to repeat or adjust. However, it also has some limits. It doesn't show the full meaning of words in context. It also treats each year the same, even if some years had more articles than others, which can affect the results. And looking at single words can miss the meaning of phrases.

Zia Ullah

TF-IDF

TF–IDF (term frequency–inverse document frequency) is a term weighting scheme commonly used to represent textual documents as vectors for classification, clustering, visualization, retrieval, etc." ("TF–IDF," 2011). To investigate how similar articles are over time, I used the TF–IDF method to generate a pairwise similarity dataset. TF–IDF is a widely used technique in natural language processing that evaluates how important a word is to a document within a corpus. It works by multiplying the frequency of a word in a document (term frequency) by the inverse document frequency, which reduces the weight of words that are common across many texts. This method, introduced by Karen Spärck Jones (1972), enabled me to represent each article as a numerical vector and compare them using cosine similarity, a score between 0 and 1, where higher values indicate greater similarity in content.

The initial dataset consisted of news articles published between 2021 and 2024. I filtered the dataset to include only articles with a minimum length of 200 words, assuming that shorter articles might lack sufficient context for meaningful comparison. This filtering yielded a more focused and manageable dataset, making it easier to draw reliable conclusions. Using the TF–IDF matrix, I calculated cosine similarity scores for each pair of articles, generating a numerical measure of content overlap. These scores formed the basis of two main visualizations: a histogram and a heatmap.

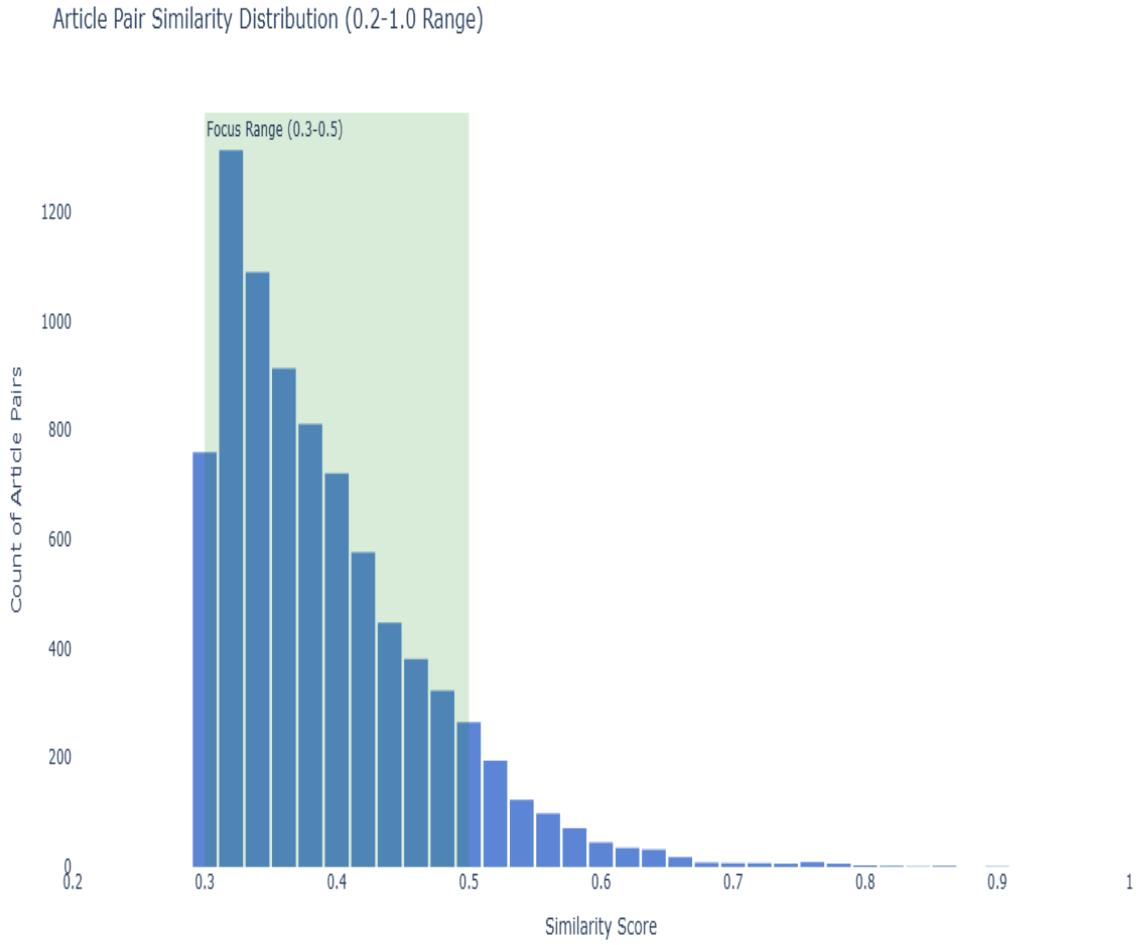
The histogram plotted similarity scores (ranging from 0.3 to 0.9) against the number of article pairs. This visualization helped me understand the general distribution of similarity within the dataset. Most article pairs had low similarity scores, particularly between 0.3 and 0.5. There was a sharp decline in the number of articles as similarity increased. Scores of 0.6 and above were rare, and scores above 0.8 appeared only once—or not at all. This suggested that strong similarity between articles was uncommon. Initially, I expected to find repetition across news cycles. However, the histogram revealed that even when topics overlapped, articles were typically written with distinct language and framing, likely due to editorial practices or the fast-moving nature of news.

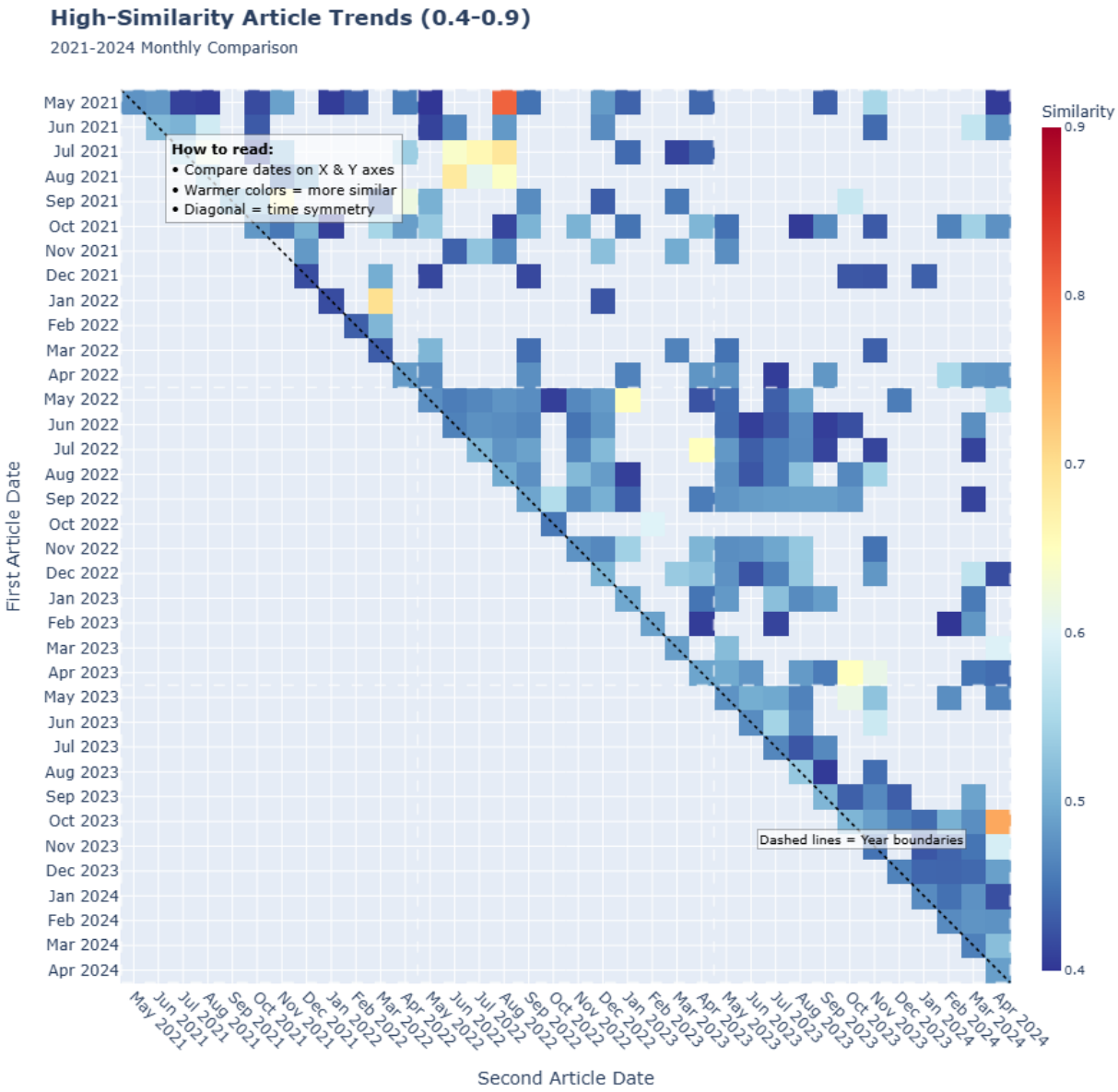
To explore how similarity changed over time, I created a heatmap that visualized the distribution of similarity scores between article pairs across different months. My first attempt included all articles and the full range of scores, but the visualization was too dense to interpret meaningfully. I refined the heatmap to include only scores between 0.4 and 0.9, which allowed clearer patterns to emerge. The improved heatmap revealed that higher similarity scores, particularly those above 0.6, clustered around specific time periods. For instance, articles scoring 0.8 appeared only once in 2021 and again in 2022. Scores of 0.7 were found sporadically, also confined to isolated months. Much of the heatmap remained empty, highlighting long stretches with no high-similarity articles.

These visualizations led me to my central argument, as similarity increases, the number of article pairs decreases sharply, and higher similarity tends to be temporally clustered. This suggests that news reporting rarely reuses identical language or framing across time. Rather, highly similar articles appear in short bursts—possibly when a significant event triggers simultaneous, similar coverage—before media outlets shift to new topics or frames. The lack of high-similarity clusters in later months also indicates that news narratives are not typically recycled.

At one point, I considered visualizing the dataset as a network graph using nodes and edges for software like Gephi. However, I found that refining the heatmap was a more effective way to highlight trends relevant to my argument. While TF–IDF has its limitations—for example, it cannot detect synonyms or semantic nuance—it was adequate for identifying broad patterns of repetition and differentiation. More advanced methods such as BERT or word embeddings could offer deeper insights into semantic similarity, but for the purposes of this analysis, TF–IDF provided a transparent and interpretable framework.

In conclusion, this exploration of TF–IDF-based similarity patterns in news articles reveals that content overlap in the media is rare and often concentrated within narrow timeframes. These findings suggest that news production emphasizes novelty and temporal specificity rather than repetition. The method allowed me to uncover both the semantic variation and the episodic nature of article similarity, ultimately supporting the argument that news is not only about *what* is said, but also *how* and *when* it is said—and rarely in the same way twice.





Reference

"TF–IDF." 2011. In *Encyclopedia of Machine Learning*, 986–87. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_832.

Arslan Danish

Topic Modeling

Topic Modeling Analysis: Identifying Latent Themes in News Coverage

To find latent patterns and recurring topics in the data, I use topic modeling to examine a big corpus of news stories in this part. In the fields of digital humanities and natural language processing (NLP), topic modeling is a type of unsupervised machine learning. By considering text as a probabilistic combination of subjects, it makes it possible to uncover hidden semantic patterns. Latent Dirichlet Allocation (LDA), first proposed by Blei, Ng, and Jordan (2003), is the method I used. It assumes that every text has several topics, and that each subject is described by a distribution of words. Because LDA can handle enormous datasets and provide interpretable outputs in the form of topics and their most likely keywords, it is frequently employed for computational text analysis.

The dataset I worked with had already undergone preprocessing and topic modeling using LDA. Each article was assigned a topic number along with associated keywords (e.g., topic_0 to topic_4). My task was to explore the structure of these topics, identify the most prominent ones, and create visualizations to interpret the data and support the report's analytical aims.

My goal was to understand how media narratives were shaped during the conflict, focusing not just on topic frequency but on their meaning, evolution, and thematic balance. I argue that media coverage tends to prioritize state-centric and security-focused narratives, while humanitarian issues receive less consistent attention. This guided my grouping of topics into broader themes and informed the visualizations used to track shifting discourse over time.

Manual Grouping and Classification:

I manually reclassified the topic modeling results into three high-level discourse topics after the exploration phase. This stage enabled a more interpretive, qualitative framing of the issues and went beyond the automatic outputs of LDA. As the previous image showed, the grouping was based on the semantic meaning of the topic terms.
I found three topic categories, which were:
1. Safety and Conflict:
Keywords like "Hamas," "military," "attack," "Hezbollah," and "missile" that were associated with war, violence, armed organizations, and geopolitical tension were included in this category. These subjects represented the fundamental media emphasis on continuing wars, military actions, and territorial conflicts, especially in areas like the West Bank, Gaza, and Lebanon.

2. Politics and Diplomacy:
This group's discussions focused on state-level participation, international diplomacy, leadership, and political negotiations. Terms like "peace," "Netanyahu," "Biden," "negotiation," and "UN" denoted discussions on diplomatic resolutions, foreign affairs, and ceasefires. These subjects

mirrored how state actors and institutional reactions to the crisis were covered by the media.

3. Aid, Crisis, and Civilian Toll:

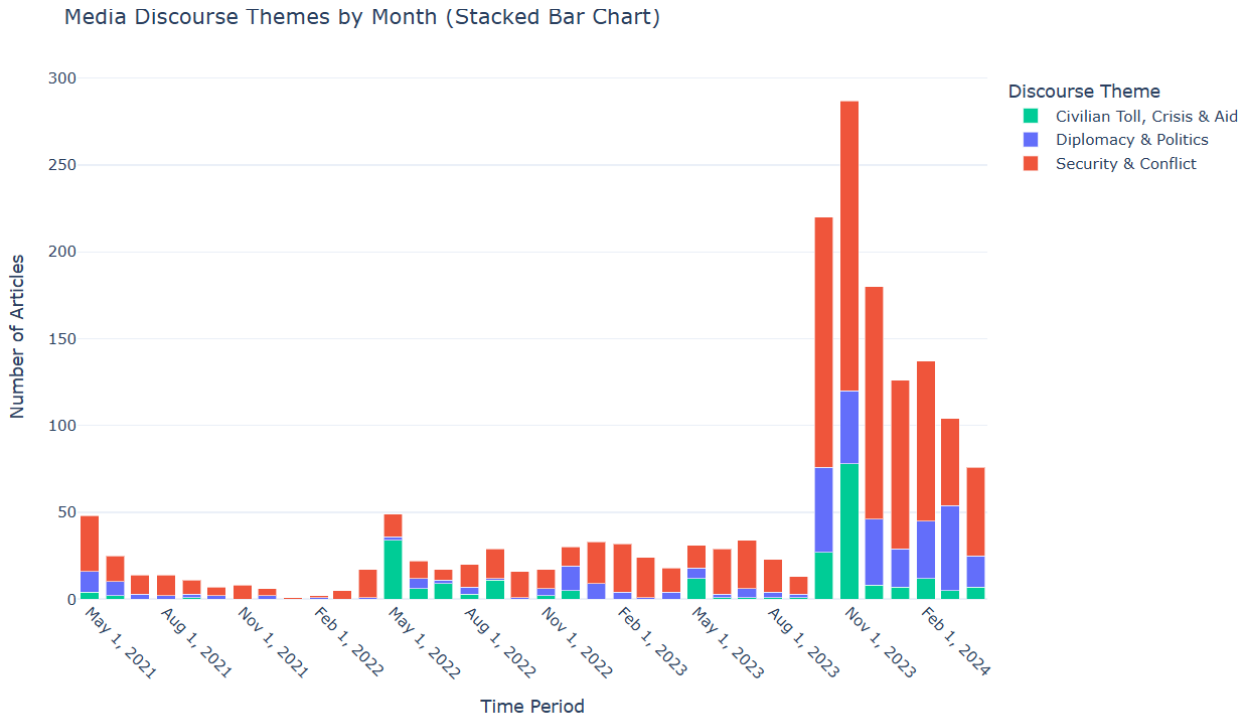
Stories about the conflict's human costs, such as civilian casualties, humanitarian emergencies, medical difficulties, and refugee displacement, were included in this category. The terms "hospital," "civilian," "aid," "child," and "hostages" were frequently used. By highlighting the suffering of citizens and the crisis response, these subjects offered a counternarrative to security discussions.

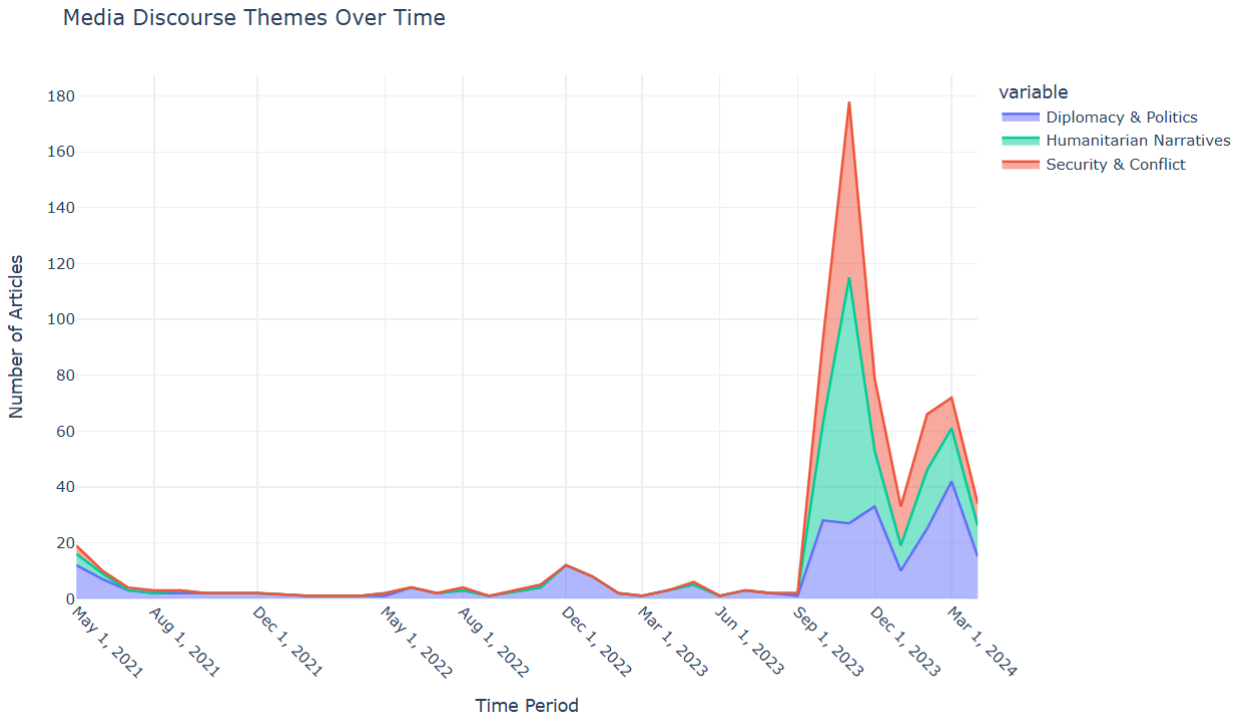
Those Articles that did not match any of the defined themes were labeled as “Other” and excluded from the final analysis to maintain thematic coherence. This manual reclassification offered a layer of critical interpretation on top of the machine-generated outputs and allowed the topic model to be reoriented around the broad interests of the project.

Presentation Visualization: Temporal Distribution of Thematic Coverage:

To show the monthly distribution of articles among the three identified themes, I made a stacked bar chart for the last step. The sections of the bar indicated the number of articles allocated to each discourse theme, and each bar represented a month. An aggregated temporal view of the movement of media attention between several narrative frames was presented by this visualization. I also made a stacked area graph to compare it with a bar graph, but the result was the same. I chose these graphs because they are very easy to interpret and are very clear.

The graphic made it evident that themes related to security and conflict dominated throughout every month of the period, particularly during times of increased military activity or escalation. Coverage of diplomacy and politics increased noticeably throughout time, especially from October 2023 onwards, sometimes in connection with international summits, peace negotiations, or policy choices. Meanwhile, attention of Civilian Toll, Crisis & Aid rose in response to specific humanitarian events, such as bombings of hospitals or refugee crises. In comparison, two other humanitarian events were less covered.



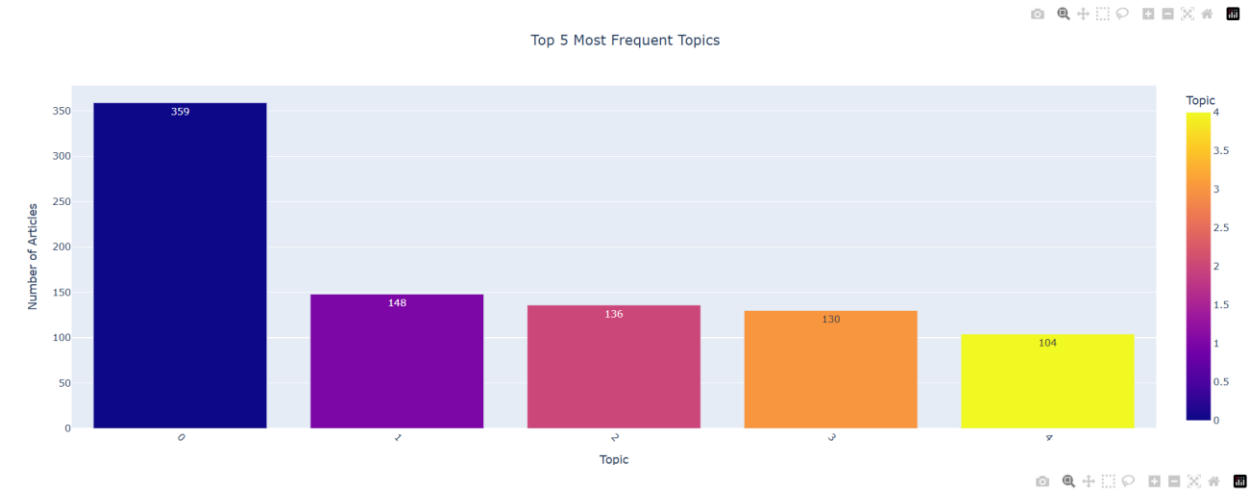


This graphic was particularly useful since it helped me make links between media discourse and actual occurrences and made the change over time visually apparent. It also emphasized how some topics were outnumbered by others, such as how coverage of civilian suffering was less consistent than that of military warfare. This supports one of our report's more general claims,

which is that media narratives frequently give priority to discourses that are state-centric and security-oriented above humanitarian ones.

Exploratory Visualization: Top 5 Most Frequent Topics and Top 5 Key Words:

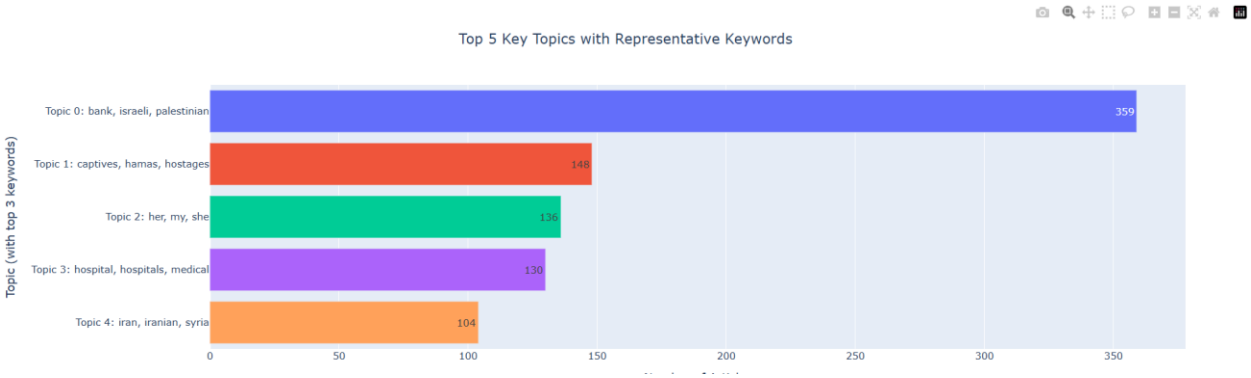
Evaluating the dataset's topic distribution was the initial stage of my analysis. I displayed the top 5 most common topics throughout the full corpus and the number of articles in each topic using a bar chart. This was a quick summary of the main themes that were covered in the news. These five subjects were ranked by article count in a bar chart, which made it simple to see which topics the model assigned most frequently. However, this output lacked semantic clarity because the subjects were initially only designated by number (e.g., Topic 0, Topic 3).



I made a second bar chart with the top five keywords linked to each of these five prominent subjects to overcome this constraint. By presenting topic labels in this visualization, "Top 5 Key Topics with Representative Keywords," it is possible to gain more understandable information. It was simpler to understand the types of discussions that formed each topic thanks to these human-readable labels.

However, I ran into a major issue when making this keyword-based visualization: several of the keywords the model produced were frequent, unclear, or semantically nonsensical (e.g., “her”,

“my”, “she”). The topic separation is not greatly improved by these so-called stop words. Despite their statistically frequent appearance, they frequently mask the underlying ideas and impair interpretability.



I created a polished script for the final presentation visualization after realizing this. I added a custom stop word filter with generic or semantically meaningless terms to this script. Articles with only stop words in their topic keywords were not included in the analysis. This stage was essential to guaranteeing the final visualization's accuracy and integrity.

Critical Reflection on the Method:

Although LDA offered a practical and scalable method for grouping texts into themes, it needs to be used carefully. Without a better contextual understanding, dependence on word co-occurrence can lead to themes that are overlapping or incoherent. Furthermore, tone, mood, and ideological setup, all essential components of media analysis, are not captured by topic modeling. Because of this restriction, to provide useful data, I had to employ manual classification, keyword filtering, and visual inspection.

Despite these limitations, topic modeling proved to be an effective exploration technique for this project. Together with analytical selection and visualization, it enabled me to map the outlines of media discourse at scale and allowed a greater understanding of the ways in which narratives about conflict, diplomacy, and humanitarianism circulate in the media.

Conclusion:

This study included Topic Modeling, TF-IDF similarity analysis, n-gram frequency tracking, and article length analysis to analyze how Al Jazeera's coverage of Gaza changed over time. These techniques showed the structural and thematic responses of media narratives to real-life events. Years like 2021 had fewer but longer articles, according to an examination of article length, indicating detailed coverage during times of increased conflict. This is consistent with the findings of topic modeling, which showed that topics related to security and conflict predominated while humanitarian issues showed up less frequently. Al Jazeera rarely repeats information, according to TF-IDF analysis, which also highlighted the episodic structure of media narratives by clustering high-similarity articles around significant events. By identifying

commonly used terms like "Gaza," "Israel," and "Palestinian," as well as phrases like "humanitarian aid," N-gram analysis provided a linguistic element and pointed to repeated frames in the coverage. All these results point to a continuous preference for conflict-driven and state-centric reporting over long-term focus on civilian suffering.

In conclusion, this multi-method approach shows how computational techniques can reveal the narratives that are recounted. It also shows the frequency, depth, and narrative viewpoints of those stories. It offers a methodology that can be reproduced for examining media discourse in various crises.