# Mini Project 3 Report

**Members:** Sara Baig (topic model), Anisha Ali (N-grams), Faiz Ahmed (Title Extraction)

**Introduction**  
For the final project for the Digital Humanities module, our group, comprising of members Anisha Ali, Sara Baig and Faiz Ahmed, was given the task to visual data from key data frames, centered around a corpus of text from al-Jazeera covering the conflict between Israel and Palestine from the period June 2 2017 till April 18 2024. Our group decided to center our argument around the use of language throughout the coverage of the conflict aiming to understand how media attention and emotional framing evolved in response to key events in the conflict. In addition we decided to hone in on data between Sep 7 2023 till April 18 2024 in order to analyse data during the October 7th attacks and the subsequent genocide that enslewed; with month long span between Sep 7 2023 and Oct 7 2023 acting as control group to compare coverage before and after October 7th 2023. Our analysis integrates three main approaches: topic modelling; utilizing the topic-modeling.csv file, article metadata analysis via the ‘lengths’ data frames, and the use of unigram data to indicate the thematic overtones of the corpus.

Using topic modelling data, we identified the 20 most frequently assigned topic labels and selected five topic sets which related to the human impact of war, these included themes of, but were not limited to: hostage situations, medical crises, and territorial control. We then tracked these topics on a monthly basis between the given time frame, which revealed how coverage adapted in response to changing conditions on the ground, such as escalations in violence or humanitarian developments for example November 2023 had a noticeably high article counts, reflecting the immediate aftermath of the conflict’s outbreak and global focus.

Alongside, we used article metadata from lengths.csv and length-year-month.csv to analyse how publication patterns changed across time and days of the week, including variations in how many articles were published in a day across the given time period, the sum and mean of the words in these articles from a monthly perspective, and the frequency of articles published in relation to the day of the week. This structural layer helped contextualise the fluctuations in thematic focus – with the largest surge in articles reported on 13 October, indicating heightened media interest in response to the aftermath of the Oct 7 attacks.

Finally, using unigram frequency data, we examined how articles related to fear and hope appeared in articles throughout the same period. Keywords such as “fear,” “grief,” “hope,” and “faith,” were the main targets of our investigation. However, our list was later expanded to include related terms. Furthermore, this data was then used to create grouped bar charts to compare monthly frequencies and was later adjusted for average mentions per article. This revealed important emotional shifts in coverage, as certain months showed a stronger presence of fear-driven language, while others reflected more hopeful or resilient tones – often correlating with changes on the ground or political developments. For example, fear-related language peaked around October 2023, which was reflecting the intensity of violence during that period, while hope-related terms saw a slight rise in early 2024, possibly connected to reports of international aid or temporary ceasefires.

Using these synergistic approaches, our group was able to highlight a more holistic analysis of al-Jazeera’s coverage which then further helped reflect the complexity of the war's humanitarian, territorial, and psychological dimensions. Through this, we argue that the media outlets not only document the conflicts but also actively shape public understanding of their changing nature by selectively amplifying themes and emotional cues at different moments in time.

**Steps and Process**

Faiz focused on analyzing how the volume and depth of news coverage on the Gaza conflict changed over time. He worked with two datasets: one with individual article lengths (length.csv) and another summarizing monthly trends (length-year-month.csv). Using Python and pandas, he extracted the publication date of each article and broke it down into year, month, and day. He calculated article length using either word or sentence counts, filtered out articles shorter than 100 words, and limited the timeframe to those published from 7 September 2023 onward. From this, he created a daily dataset and then grouped it monthly to calculate the total and average article length using aggregation methods like sum and mean. He visualized daily article frequency using a Plotly bar chart, which showed a spike after the conflict began on 7 October 2023, with the highest number of articles published on 13 October. Faiz also examined which weekdays had the most articles, finding that Thursdays and Fridays saw the most activity. Finally, he used the monthly dataset to assess how article lengths changed over time, finding that November 2023 had both the longest articles on average and the highest overall word count. All these visualisations were then subsequently saved as HTML files. His process demonstrated how filtering by word count and analyzing trends in frequency and length can help track media coverage frequency.

Anisha was responsible for working with the 1-gram-year-month dataset to analyze emotional tone in the articles using distant reading. Her goal was to compare the use of fear-related and hope-related words over time and see how emotional language evolved across the Gaza conflict. She started by exploring the dataset to understand its structure, including fields like year, month, 1-gram, count-sum, and count-mean. This helped her see what kinds of frequency comparisons were possible. She then created an initial script of exploration to filter a basic list of emotional terms like “fear,” “grief,” “hope,” and “faith” and tested how often they appeared over time using a line graph. This exploratory step helped her understand how the dataset was organized and grouped.

She then created more visual scripts. In her first main visualization, she expanded the word lists to include terms like “panic,” “death,” “resistance,” and “courage,” grouping them into two categories: fear and hope. She used absolute frequency (count-sum) to make a grouped bar chart comparing monthly totals of each category. After feedback from the instructor, she improved her approach by including word variants (like “hopes,” “hoping,” “died,” “deaths”) and aligning her timeline with her groupmates from 7 September 2023 to 16 April 2024. In her final script, she switched from raw counts to relative frequency by using the count-mean column to show average mentions per article. This made the data easier to compare across months with different article volumes. Each of her scripts was visualised and exported as HTML. Her process showed how emotional tone in media can be tracked over time, and how refining word lists and metrics leads to better insights.

Sara worked on analysing the key themes/topics that appeared in news articles between 7 September 2023 and 16 April 2024. She used a dataset called topic-model.csv, which listed articles along with a basic topic number and four keywords that described each one. She split her work into two scripts: one to explore the data and one to visualise the results.

In the exploration stage, she filtered the data to focus only on articles from the group’s chosen timeline. She then looked at the top 20 most common topics and their keywords to understand what each one was about. One of the topics included only pronouns, which didn’t provide useful information, so she replaced it manually with a clearer, more relevant one related to the conflict.

To track changes over time, she created a month\_year column and combined the keyword columns into a single, readable topic label. She grouped the data by these labels and month to see how often each topic showed up. An initial chart of the top 20 topics helped guide the selection for the final visualisation.

In the visualisation script, Sara focused on just five meaningful topics. She created a grouped bar chart to show how frequently each topic appeared from month to month. This helped highlight how the media’s focus changed during the war, and supported the group’s overall argument of understanding the effects of the conflict through patterns in news coverage.

**Visualisations and Findings**

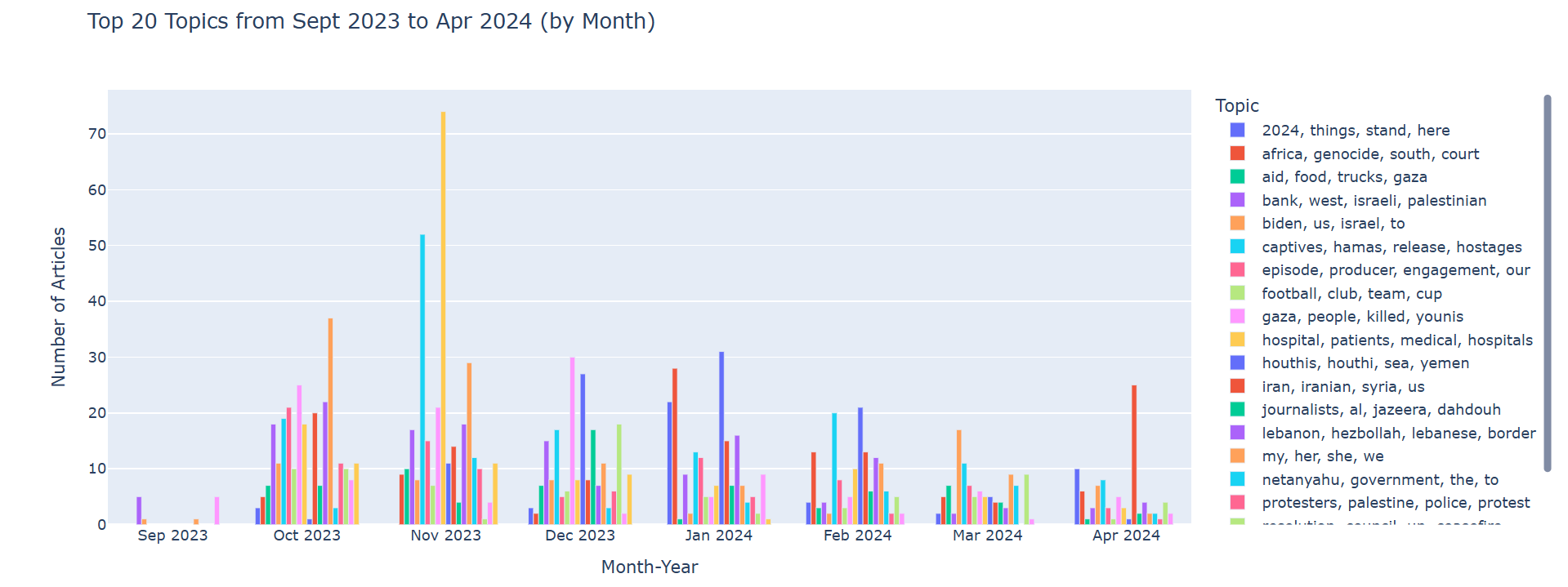
1. **Topic Modeling by Sara Baig:**

For this project, I worked with the topic-model.csv dataset, which was produced using Latent Dirichlet Allocation (LDA), a popular unsupervised machine learning method for topic modelling. This dataset included article-level metadata from the al-Jazeera Gaza corpus, consisting of columns such as date (year, month, day), article title and four keywords (topic\_1 to topic\_4) representing the dominant topic of each article. LDA assumes that documents are mixtures of topics, and each topic is itself a mixture of words. In other words, it probabilistically assigns topics to documents and identifies which words most likely belong to each topic.

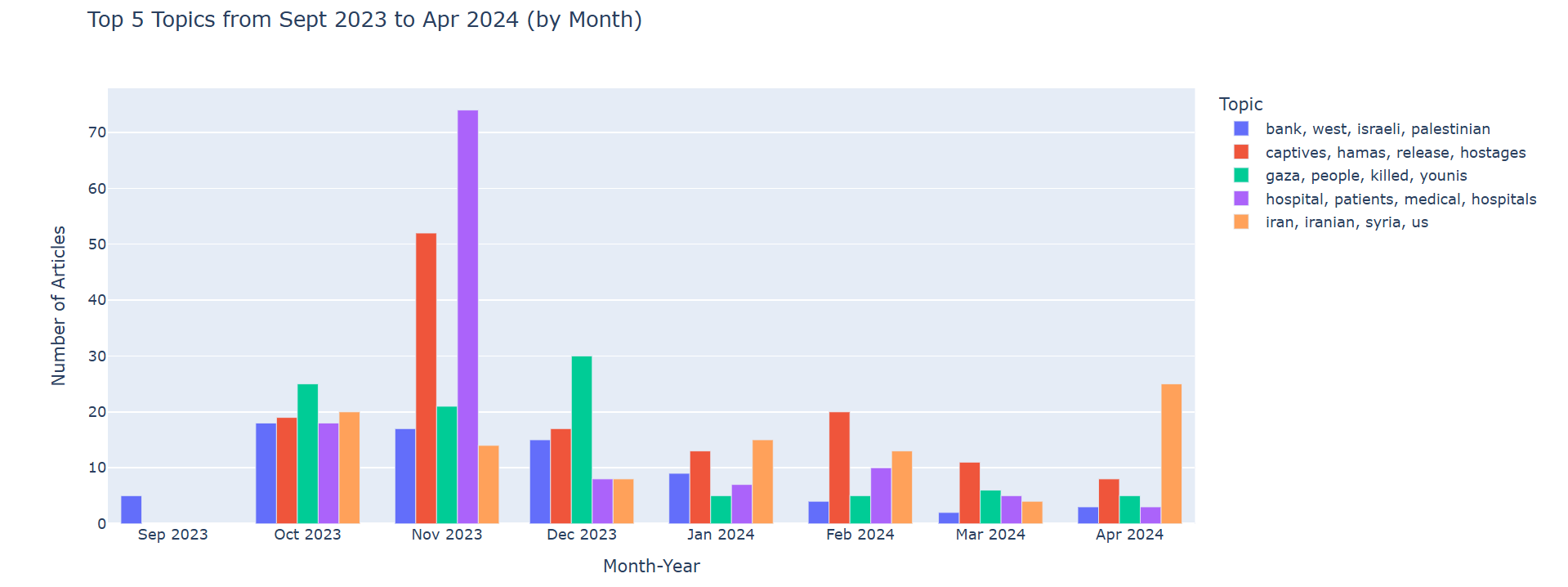
The LDA method is particularly well-suited for uncovering patterns in large text corpora, as it reduces the dimensionality of text data by grouping related terms under hidden topics. Each document in the corpus is thus represented not by individual word counts but by its association with latent topics, making it easier to interpret broad thematic trends (Kapadia 2025) Our project aimed to understand how Al Jazeera’s coverage of the Gaza conflict evolved in response to key events. LDA was an appropriate choice because it allowed us to track recurring, event-driven themes across time.

LDA is not without limitations. It requires pre-selecting the number of topics, and if poorly tuned or insufficiently preprocessed, it can generate vague results. This was evident in our case when one of the top five topics included only personal pronouns like “my,” “her,” “she” and “we.” As these were not thematically meaningful in the context of war coverage, I removed this topic manually and replaced it with a clearer one focused on territorial terms: “bank,” “west,” “Israeli,” and “Palestinian.”

My topic modelling analysis included two graphs. The first, generated during the exploration phase, displayed the top 20 most frequent topics within the timeframe of 7 September 2023 to 16 April 2024. Each bar represented a topic label (combined from topic\_1 to topic\_4) and showed its frequency over time. This visualisation allowed me to assess which topics were most dominant and helped me in my final selection of five conflict-relevant topics.

****

The second visualisation generated during visualization phase focused on those five chosen topics. The grouped bar chart showed how frequently each theme appeared month by month. Several key insights emerged: November 2023 saw a spike in article counts across all selected topics, corresponding to heightened global focus following the war’s escalation. Topics related to casualties, hospitals, and Gaza-based humanitarian crises showed pronounced peaks during periods of major bombardment, while hostage-related and geopolitical topics like Iranian/Syrian involvement were more steady but still significant. The inclusion of the territorial topic added another layer, showing that land-related discourse was present but not dominant.



These visual trends emphasized that media coverage shifted thematically and emotionally in response to conflict developments. The topic modelling visualisations showed that Al Jazeera’s reporting evolved with key humanitarian and political events, providing strong evidence that journalistic focus adapts in real time to reflect on-the-ground realities. In this way, LDA helped us meaningfully interpret how the conflict was framed and followed through media narratives.

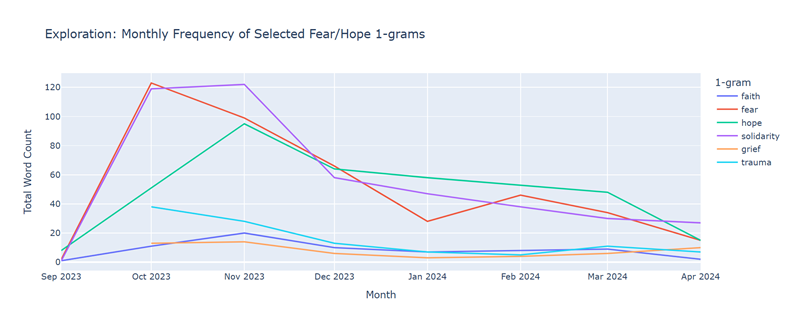
1. **N-grams by Anisha Ali:**

My part of the project was related to the n-grams where my goal was to study how language in the media reflected fear and hope during this time. I wanted to see if fear-related terms increased during violent events and whether hope-related terms appeared during calls for peace. I compared the use of fear-related and hope-related words in the news articles between September 2023 and April 2024.

I worked with the 1-gram-year-month dataset, which includes the monthly frequency of individual words (1-grams) used in the Al-Jazeera Gaza corpus. This dataset has word counts as well as average mentions per article (count-sum and count-mean). It also has date markers such as year and month. This dataset was produced by using a simple and widely used natural language processing method called tokenization. Tokenization breaks down the corpus into individual words. These words are then counted across documents and grouped by time (Šišnović 2024). This method is commonly used in digital humanities and large-scale text analysis, such as Google’s N-gram project, to detect historical or emotional patterns in language over time.

So, at first, I started by exploring the structure of the 1-gram-year-month dataset. My goal here was to see how often fear-related and hope-related words were used in Al-Jazeera’s coverage of the Gaza conflict. I created a basic word set for each category. For fear, I selected “fear,” “grief,” and “trauma,” and for hope, I selected “hope,” “solidarity,” and “faith.” These were the most direct emotional indicators I could think of.

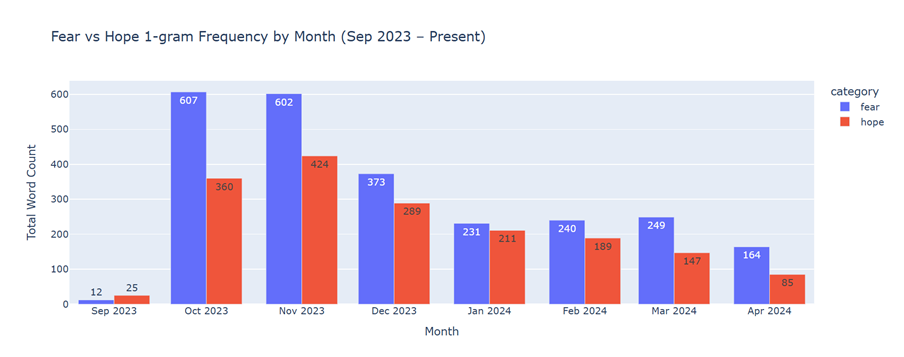
In my exploration script, I plotted a simple line chart to visualize the frequency of these words over time. This helped me see which words were being used the most and in which months. After that initial exploration, I moved on to a second script where I expanded the word lists to make the analysis more meaningful. I added "panic" and "death" to the fear category, and "resistance" and "courage" to the hope category. This gave a richer emotional spectrum. I then created a grouped bar chart to compare the absolute frequencies (count-sum) of fear vs hope words across months. While this method was useful for getting an overview, I realized that raw counts could be misleading, especially when the number of articles published each month varied. This led me to critically rethink the metric I was using.



A graph of different colored bars

AI-generated content may be incorrect.

After taking some feedback from my professor into account, I began my third script. I extended my word lists to include multiple variants (like "hopes," "hoping," and "fears," "died") and adjusted the timeline to match my groupmates’ work, i.e., starting from 7 September 2023 and ending on 16 April 2024. This script still used absolute counts, but it reflected a more thorough filtering process and timeline alignment.



My final visualization showed a grouped bar chart focusing on the average number of mentions per article for both fear and hope-related words from 7 September 2023 to 16 April 2024. The graph shows a noticeable spike in fear-related language in October 2023, likely tied to the outbreak of the Gaza conflict. In contrast, a modest rise in hope-related words appears in early 2024, which may reflect moments of international aid or calls for ceasefire. These trends suggest that media coverage does not only report events but also frames them emotionally using language that may influence how the public understands and reacts to the conflict. Finally, this emphasizes that emotional language in media shifts with the intensity and nature of events and helps shape public perception through subtle linguistic choices.

A graph of red and blue bars

AI-generated content may be incorrect.

1. **Title Extraction by Faiz Ahmed:**

In this section, I explored how news outlets covered the Gaza conflict over time by looking at how many articles were published each day and how long those articles were. I used two main datasets: one with individual article lengths (length.csv) and another that summarized monthly trends (length-year-month.csv). To generate these data frames, each article in the corpus was looped through, and its publication date is divided into year, month, and day using panda's datetime module. The article content was then processed to calculate its length, typically by splitting the text into words (split()) or tokenizing into sentences (e.g., with nltk.sent\_tokenize()). This information was stored in a structured list and converted into the lengths.csv data frame using pandas. For the length-month-year.csv, the initial data was grouped using groupby(['year', 'month']), and aggregation functions like sum and mean are applied to the length column (McKiney, 2018). This transforms detailed article data into a compact summary of content volume and average length per month.

Additionally, I filtered articles that were shorter than 100 words were filtered out to ensure that the analysis focused only on articles with substantial content (WordCount, 2017). Furthermore the data was filtered to include only articles published from September 7, 2023 onwards, covering both the period before and after the escalation of conflict on October 7. Lastly python's libraries, specifically the Pandas and Plotly were used to process and visualize the data.

Figure 1 below shows how many articles were published each day. Articles were grouped by date and a bar chart was created using Plotly. As expected, there's a huge spike in coverage right after the conflict began. October 13, 2023 had the most articles, with 34 published that day. This could reflect al-Jazeera's response to the escalating situation during this period.

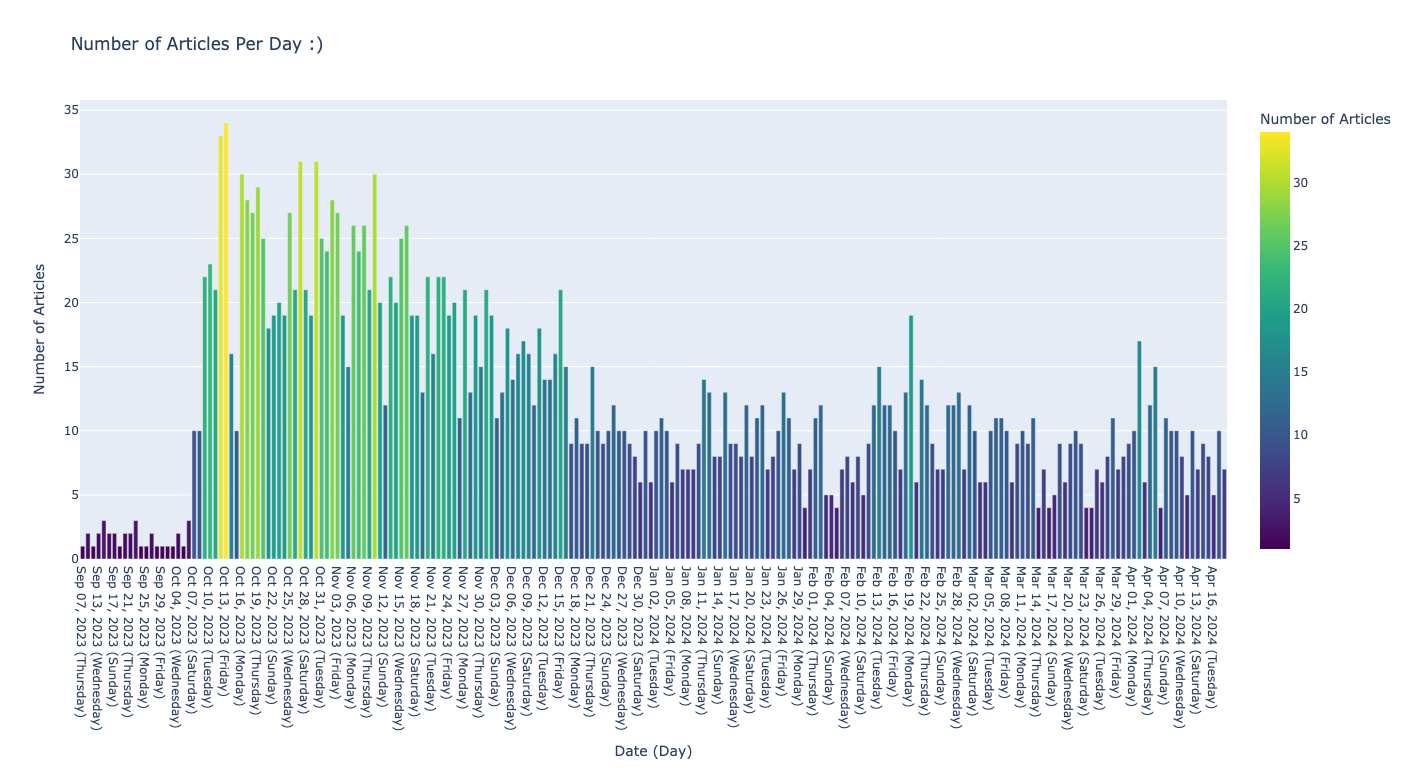


Fig 1.

Next, I looked at which days of the week had the most coverage. By extracting the weekday from each date, he found that Thursdays and Fridays had the highest publication rates, while Sundays had the lowest. This likely reflects newsroom work cycles and publishing routines (Zlatin, 2025).

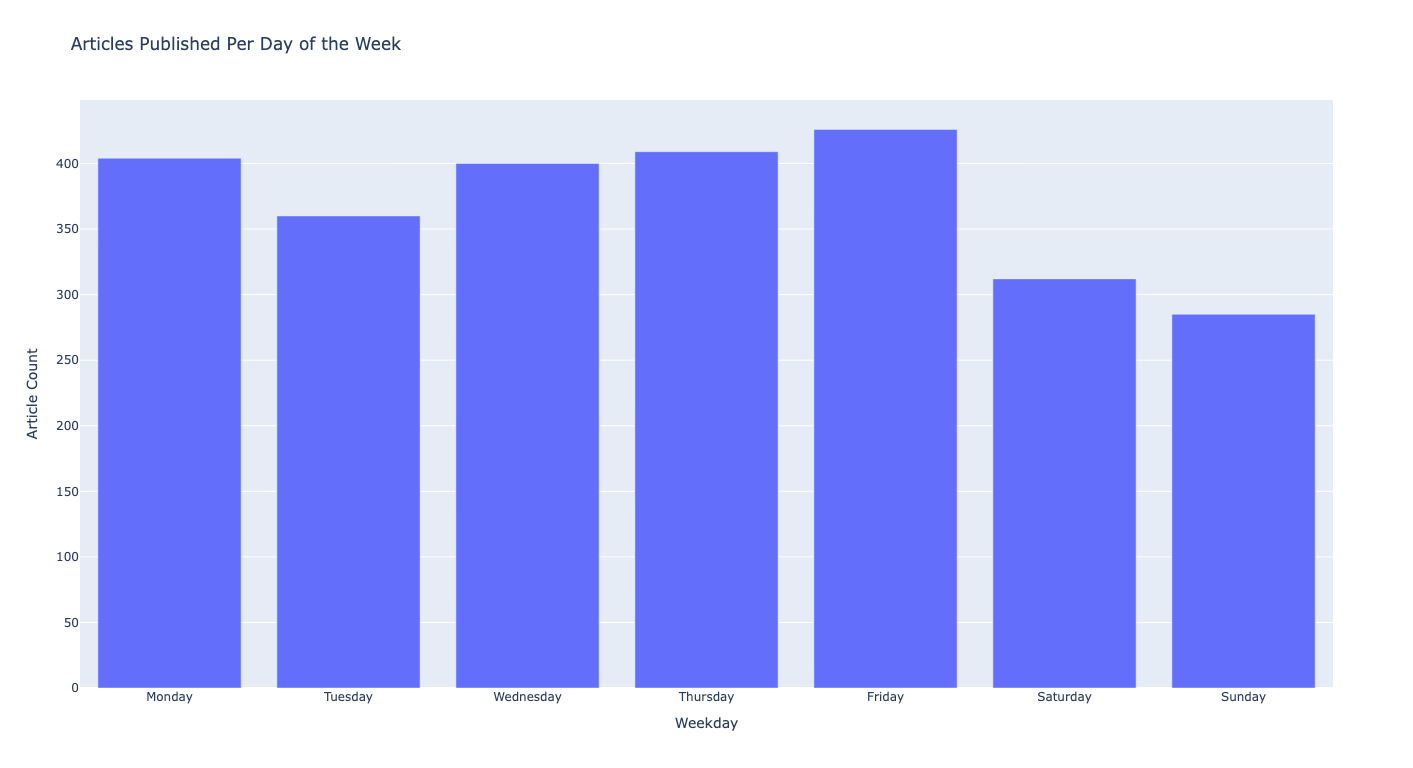


Fig 2

Finally, the change in article length was analyzed on a monthly basis, taking into account both mean length as the sum of the words for each month. Data was accumulated from the second dataset, which utilised the length-year-month.csv. The results indicated that November 2023 had the longest articles on average with a mean of around 649.39 words. This aligns with journalistic standards, as early reporting often includes background and context (Zelizer and Tenenboim-Weinblatt, 2014). In later months, like April 2024, mean article lengths dropped to about 596.34 words, likely because the news had shifted to quicker updates or reporting had decreased due to a rise in other geopolitical events. Additionally the sum of the length of the articles was at its apex in Nov 2023 with a reported sum of 444,184 words, which when coupled with the average mean indicate that the most amount of articles were published during the initial stages of the conflict.

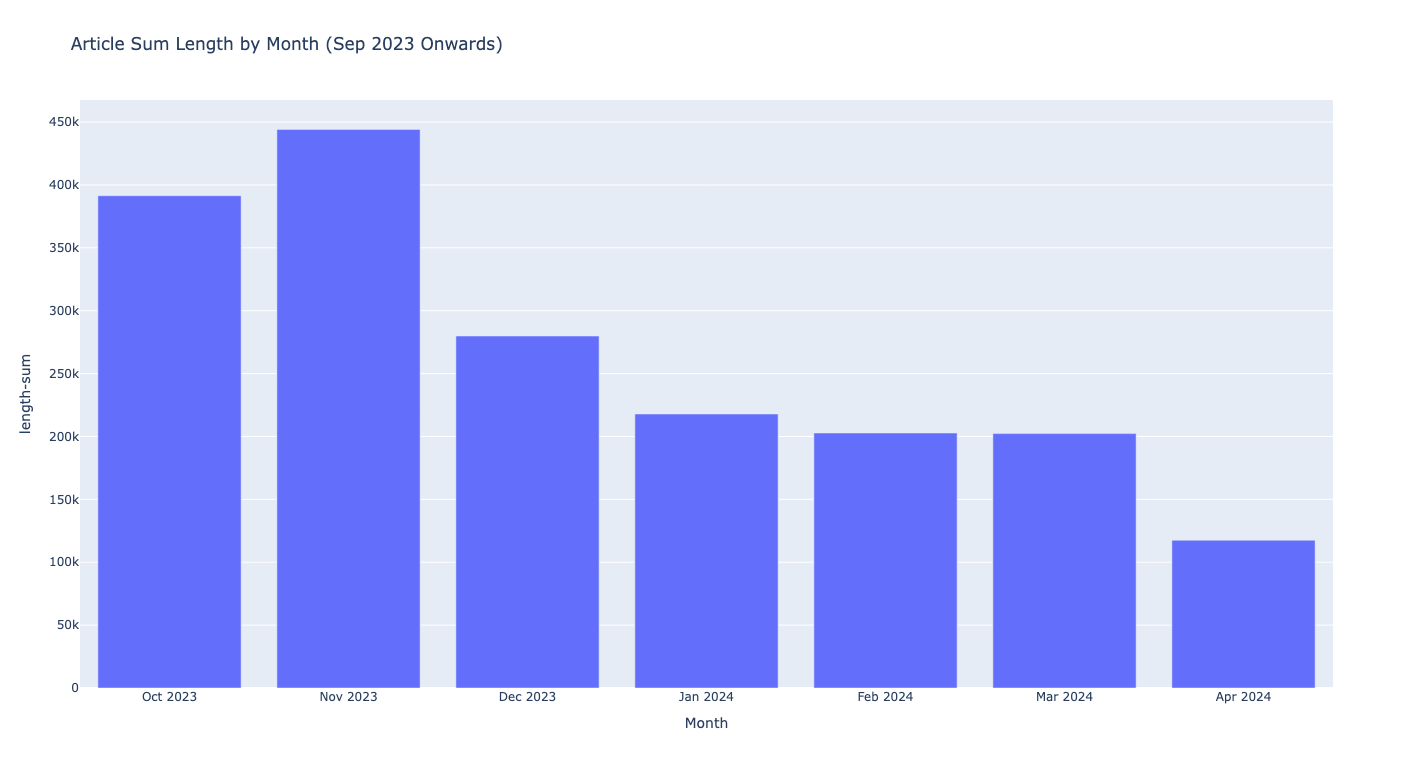
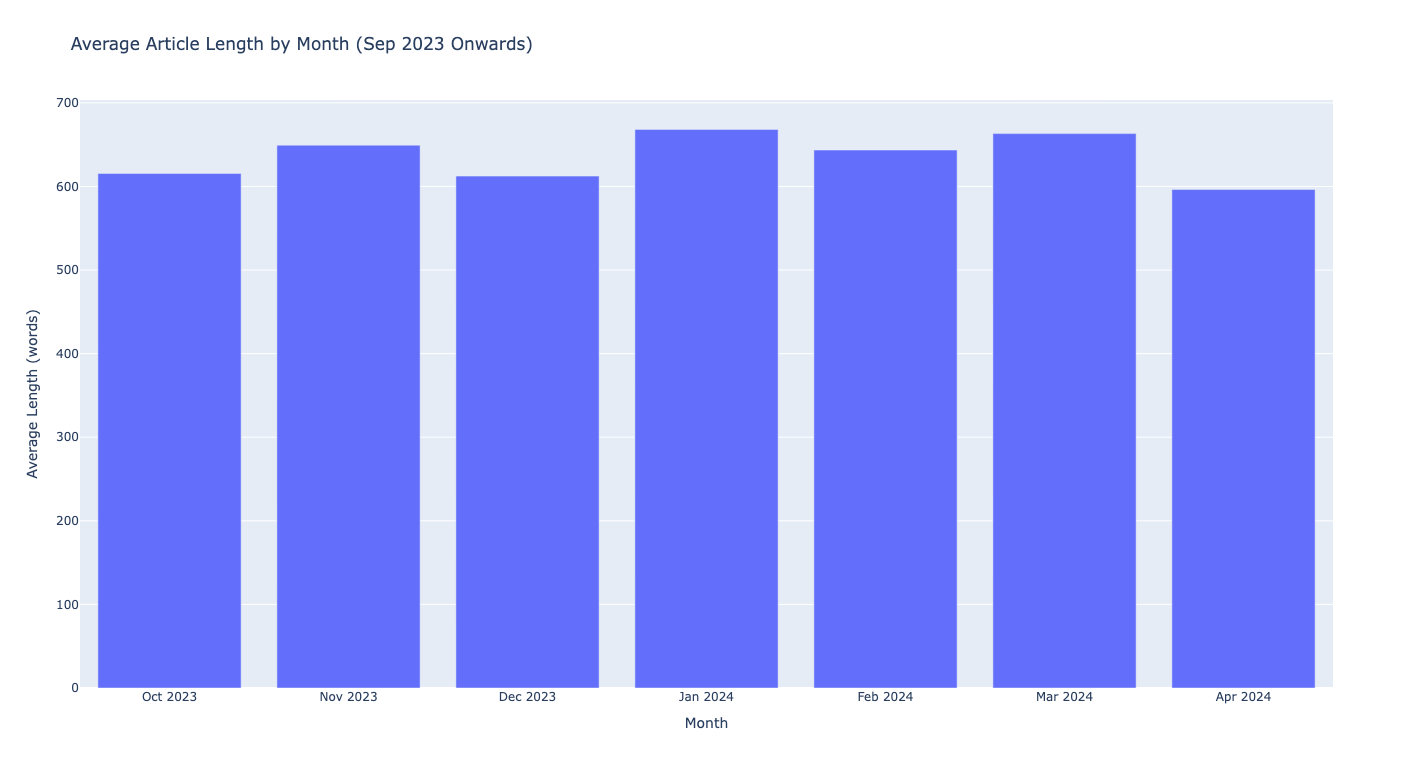


Fig 3 Fig 4

Overall, filtering by word count and looking at both frequency and length turned out to be a helpful way for the group to track how deeply and how often news outlets covered the conflict. While article length isn’t a perfect measure of quality,being unable to capture tone or nuance, it aids in providing an outlook to measure reporting frequency.

**Conclusion**Our project set out to trace how Al Jazeera’s coverage of the October 7 attack and subsequent war on Gaza evolved in content, structure, and emotional tone between September 2023 and April 2024. By combining topic modelling, article metadata, and emotional language analysis, we uncovered distinct patterns in how the conflict was reported and framed over time.

The topic modelling analysis identified five dominant and interpretable themes that structured Al Jazeera’s reporting throughout the conflict: hostage situations and international diplomacy; civilian casualties and medical infrastructure; ground operations and territorial control; humanitarian aid and ceasefire efforts; and Israeli domestic politics and military planning. Tracking these themes over time revealed how coverage responded to major events, for example, a rise in diplomacy and hostage-related content during truce negotiations, and a surge in medical crisis themes during periods of heavy bombardment. Territorial and military themes were especially prominent in October 2023 and January 2024, aligning with phases of intensified ground operations.

Our metadata analysis provided additional context. The article count histogram clearly indicated peaks in publication frequency in October 2023 and January 2024, which aligned with the outbreak of the war and major escalations respectively. The weekday bar chart indicated that most articles were published on weekdays, particularly Tuesdays and Wednesdays; this was found to be consistent with the newsroom publishing cycle. Meanwhile, the article length graph showed that articles were longest on average in October and November, suggesting a need for more in-depth coverage during the early and most chaotic stages of the conflict. Shorter average lengths in later months may reflect either reporting fatigue or streamlined updates as the war settled into prolonged phases.

The uni-grams analysis, focusing on the normalized frequencies of unigram terms associated with fear (e.g., “fear, ”“grief,” “trauma” ) and hope (e.g., “solidarity,” “faith,” “hope”), revealed how sentiment tracked closely with the war’s dynamics. Spikes in fear-related terms aligned with escalations in violence, particularly in October 2023 and January 2024, when ground operations intensified and casualty reports surged. Hope-related language, while generally less frequent, peaked during brief periods of truce or humanitarian aid efforts, most notably during the November ceasefire negotiations. Importantly, normalizing these emotions by article count exposed that while the raw frequency of fear terms dominated, moments of diplomatic progress saw the highest per-article intensity of hopeful language, indicating that emotional framing was not solely driven by article volume but by editorial emphasis on sentiment during pivotal periods.

Together, these findings support our argument that Al Jazeera’s reporting evolved in line with both the material developments of the conflict and the emotional landscape it provoked. Thematically, structurally, and emotionally, the data reflect a media narrative that responded dynamically to unfolding events, helping shape how the war was understood and felt by its audience. The interplay between fear and hope in the language of reporting adds a nuanced layer, showing how sentiment mirrored both devastation and resilience over time.

**References**

Kapadia, S. (2019, April 14). *Topic Modeling in Python: Latent Dirichlet Allocation (LDA)*. Towards Data Science. <https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0/>

McKiney, W. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.)*. O’Reilly Media.

Word Counter. (2017, April 4). Word Count List – How Many Words Your Text Should Have. *Word Counter Blog*. <https://wordcounter.net/blog/2017/04/04/102966_word-count-list-how-many-words.html>

Zelizer, B., & Tenenboim-Weinblatt, K. (2014). Journalism’s Memory Work. In B. Zelizer & K. Tenenboim-Weinblatt (Eds.), *Journalism and Memory* (pp. 1–14). Palgrave Macmillan UK. <https://doi.org/10.1057/9781137263940_1>

Zlatin, G. (2025). *How to Create Content to Tap Into the News Cycle*. <https://www.digitalthirdcoast.com/blog/how-to-tap-into-the-news-cycle-with-your-conten>

Šišnović, Matej. 2024. “Tokenizing Sentences and Datasets in Python (for Machine Learning Purposes).” *Medium*, November 16, 2024. <https://medium.com/@m.sisnovic/tokenizing-sentences-and-datasets-in-python-for-machine-learning-purposes-4b9eba7b23ed>.