**Report**

**Introduction**

The Gaza corpus contains numerous articles particularly related to the ongoing conflict between Hamas and Israel. As a group of four members, we used four different data frames: counting the lengths of the articles, with separate tables grouping lengths by year and month (Wajahat Ali), TF-IDF cosine similarities between articles (Suleman), n-gram frequencies (1-gram, 2-gram, 3-gram) with separate tables grouping frequencies by year and month (Shehzad) and topic modelling (Sohrab). By studying and analyzing the CSV files especially title CSV, it was observed that the average title length usually changes on a yearly basis and the number of titles mentioning war related terms generally remained high and increased in recent years. This analysis illustrates that the topics about war and conflict are frequently discussed and covered in the media, remaining relevant over time.

**Documentation:**

In the mini project-3, Wajahat was responsible for forking and cloning the repository. Moreover, he also set up the Trello board, invited the group members, and distributed tasks. Wajahat worked with the length and title CSV files. Firstly, he worked with the exploration script and produced some rough outputs. Afterwards, he started working on the visualization script where he produced four different visualizations.

* articles count per year   
  articles count per month.
* Total article length per month
* War terms titles per year

To generate these graphs firstly libraries like pandas and plotly express were imported then the csv file location was uploaded with the help of these libraries. after wards and those files were read and exported using necessary codes. Libraries like pandas and plotly are very important for date manipulation, analysis and cleaning.

**Wajahat Ali: Analysis**

**Article Lengths and War-Related words (2017-2018)**

The datasets I used for my visualizations was lengh.csv and title.csv, particularly relying more on the latter. The focal point of my analysis was to investigate how has the war related language or words have been used over the years, and whether their usage increased or decreased between the 2017-2013 period.

As already mentioned in this report, to conduct this analysis, I first Imported pandas, which is an important built-in library to upload and read csv files. Secondly, I analyzed the articles count and length per year and month to check the trends. The main goal, however, was to check the war related words such as attack, killed and bombardment etc. I defined these words, ran necessary commands and generated visualization graphs to track their usage.

The lambda function with .apply() was used to check the occurrence of any of these words in the articles. In addition, function like. group by() and. sum () was used to create a data set of were related counts. Lastly, another built-in python library, plotly, was used to generate visualizations for analysis.

From figure 4 in the visualization output, it became evident that the usage of war related titles has gradually increased from 2017 until 2021.However, 2021 onwards the graph shows an abrupt surge in war related terms, which concurs with the initial argument made in the introductory paragraph in this report. The sudden increase in war terms can be related to the articles count per year, as 1700 more articles were published in 2023 as compared to 2022. This visualization also supports the argument that increase in global conflict draws more media attention, leading to rise in the use of war related words in media and articles.

A graph with a line going up

AI-generated content may be incorrect.

A graph with numbers and a bar

AI-generated content may be incorrect.

**Sulaiman Essa:**

I used a network-based strategy by displaying TF-IDF cosine similarity values in Gephi in order to investigate content similarity within the Al Jazeera Gaza corpus. Each node in this network represents an article, and edges link articles with cosine similarity scores greater than 0.9 that have very comparable content. By setting this criterion, the network is guaranteed to contain only strong links, which are an indicator of substantial textual overlap.

After importing the data into Gephi, I used layout methods like Fruchterman-Reingold to distribute the nodes according to how strongly they were connected. Two items displayed closer together in the graph the more similar they were. I used Gephi's modularity method to perform community detection in order to make the structure easier to understand. Articles were automatically divided into clusters according on how connected they were. Because each cluster was given a unique hue, it was possible to visually identify groups of articles that most likely discussed the same topics or events.

The resulting image showed a distinct structure: closely linked articles arranged in densely knit clusters, indicating that the same issue had been covered or updated. Some articles, which might have been very significant, regularly cited, or republished works, showed up as central nodes with numerous links to other articles. I colored the graph's nodes according to the month of publication in a different version, which showed that articles from the same month tended to group together, suggesting temporal consistency in reporting themes.

Overall, we saw how theme communities were created by article similarities through Gephi visualization. It brought attention to the news cycle's temporal grouping, narrative clustering, and content duplication patterns. By successfully converting abstract similarity scores into a concrete visual structure, this network-based method advanced corpus analysis beyond what was possible with conventional charts.

**Main argument:**

Together, the visualizations created by this study lend credence to the claim that the Gaza corpus is organized around many unique groups of remarkably similar articles, exhibiting trends in theme- and event-driven reporting as well as notable content repetition across time. I was able to spot distinct patterns in the way news coverage changed over the course of the conflict by analyzing the overall distribution of similarity scores using a histogram, mapping the structural relationships between articles in a network graph, and investigating the temporal dynamics of article similarity using line and scatter plots. While persistent clusters of extremely similar articles demonstrated the existence of recurrent narratives and frequent content reuse within the corpus, similarity peaks correlated with significant events. When taken as a whole, these results show that the reporting in this corpus is not dispersed equally but rather is arranged according to major plot points and times of high media attention.

A screenshot of a computer

AI-generated content may be incorrect.

**A graph showing a graph

AI-generated content may be incorrect.A graph of a bar graph

AI-generated content may be incorrect.**

**Topic Modeling Analysis (Suhrab-Wali)**

**Topic-modeling and Methodology:**

Topic modeling is an unsupervised machine learning approach that identifies hidden thematic structure in the corpus by grouping words with similar tendencies to occur together. In our work, this method was applied to uncover main topics presented by news headlines. To find the most important topics within our dataset of news articles, I used a technique called Latent Dirichlet Allocation (LDA). LDA is a topic modelling technique that discovers hidden topics in large volumes of text. It does so by comparing how often words occur in the same article, then grouping them into clusters, or “topics,” according to the patterns it finds. For instance, if the terms “hospital,” “patients,” and “medical” frequently appear together in the text, LDA cluster them into a topic about health. This approach is a popular technique in the digital humanities and social sciences for finding patterns of representation in public discourse. Using these article-specific monthly counts, I labeled the top five topics based on their most frequent keywords and monitored them over time. This shows insight into how attention to particular subjects, whether health care, military conflict or international politics, varied month by month.

We used existing CSV file (topic-model. csv) containing predefined topics, keywords, and the metadata of the articles, to complete our task. This helped us to focus on visualization and interpretation of the Data, instead of building from scratch. My part was to explore topic modeling and create two scripts and generate a series of visualizations and outputs. I created both Exploration Script and Final Visualization Script, which provide both quantitative understanding and narrative insights about our structure of our corpus. Both scripts used the “topic model” CSV data, which contained document IDs, date metadata, assigned topics, and topic-specific keywords.

**Exploration Script Outputs**

The exploration script I developed was an overview and a preparatory phase for what I will visualize at my final script. I went through different steps, explored topic modeling Data set, and created two graphs (visualizations) and two CSV files for exploration outputs. The graphs include bar chart about “Article Counts by Topic and Year”, and a Line Chart showing “Daily Topic Trend Over Time for Top 5 Topics”. While the CSV files I worked on were “War Mentions per Year” and “Article Count per Topic”. These selections were chosen as they support the overall project argument by revealing both macro and micro-level patterns in topic distribution and narrative emphasis over time. As the graph and CSV file shows it clearly that the number of ‘War’ mentioned in articles increased from 3 in 2022 to 122 in 2023. Which clearly justifies that War related words and themes have been highlighted in the media in recent years. The Daily Topic trends graph also shows that during October 2023 and February 2024, topics like Hospital, patients, captives, hostages, and Hamas had higher frequency than other topics which clearly shows that during this time war related topics were trending in these articles.

Alongside these outputs, I also filtered out all unassigned topics where (Topic = -1) to make sure that only meaningful, assigned topic data was analyzed. After that, I removed common stop words from the four topic keyword columns (topic\_1 to topic\_4). These stop words are common words like "the," "is," and "and" that don't add any real meaning to the analysis, I took help from the NLTK library to find and remove these words. I used a function on each keyword column to see if a word is a stop word and take it out if it is. This step was necessary to clean the data and make the topic keywords better match the actual content themes. This made the visualizations and keyword-based interpretations clearer. Exploration script and output generated from them helped me to understand not only code related difficulties and issues about how to make visualizations, but it also helped me to explore the Data set and know how it relates to my group’s argument.

**Final Visualization Script**

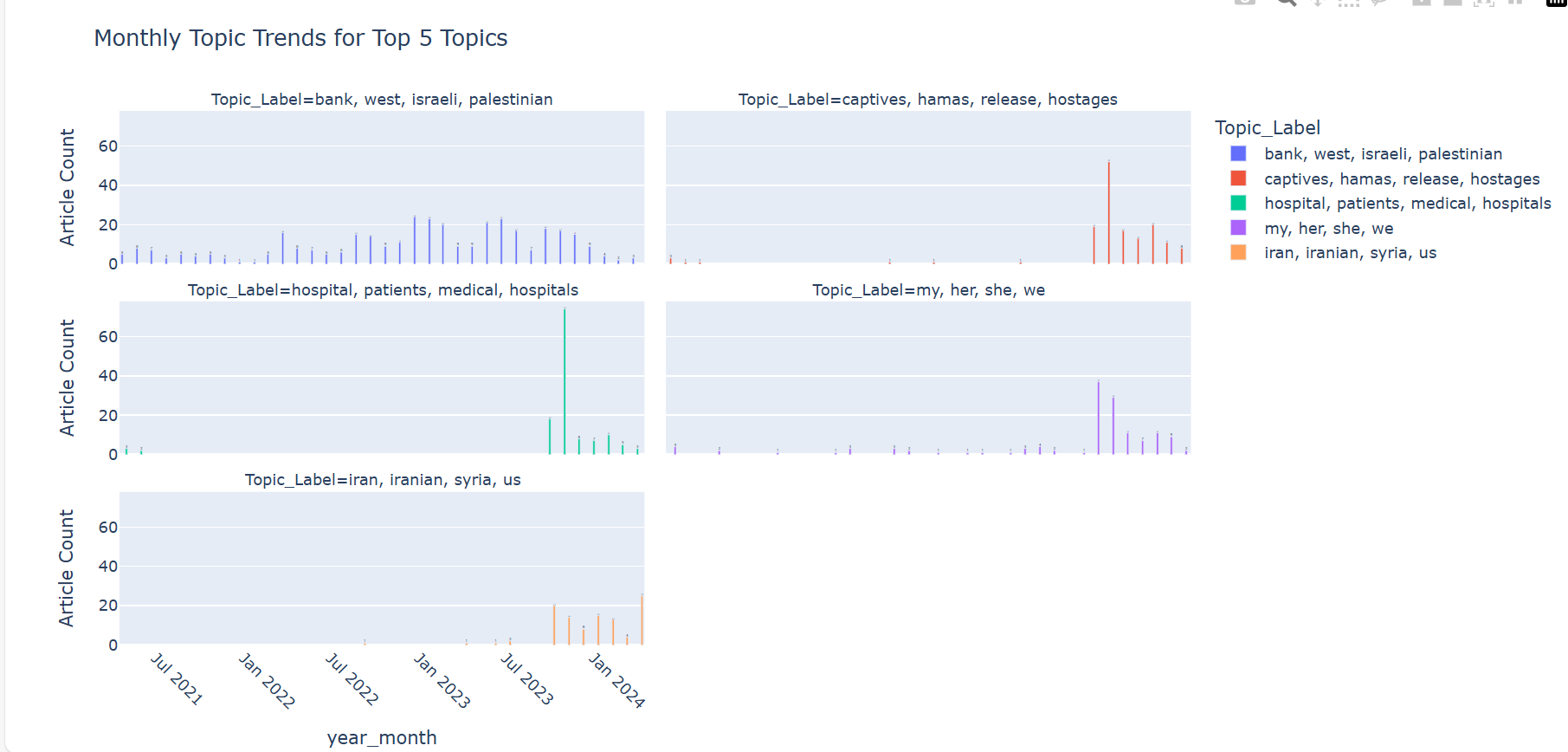
In the final script I focused on exploring the structural and over changing patterns in Gaza topic-model dataset. This script I generated was to produce four visualizations which extended and aligned with our group’s argument about how war and conflict have been regularly covered in the media and remain important over time, and how repetition of themes and narratives evolve over time. At the start, I cleaned the data by combining topic keywords and filtering out unassigned topics.

**Figure 1:** **Number of articles per topic**

Then I created first graph about “number of articles per topic”, which reveals that the highest number of articles are on topic like “Israeli forces, Palestinians, west, killing, etc.”, and at second place topics are about “captives, Hamas, release, and hostages”. So, these findings clearly show the importance of this conflict and how themes of war, Israel, and Palestine have regularly been covered in these articles.

**Figure 2: Monthly Topic Trends**

Then, to trace how these topics evolve over time, I developed a “monthly trend Visualization”, which highlights changes in topic intensity and showing spikes around key events.



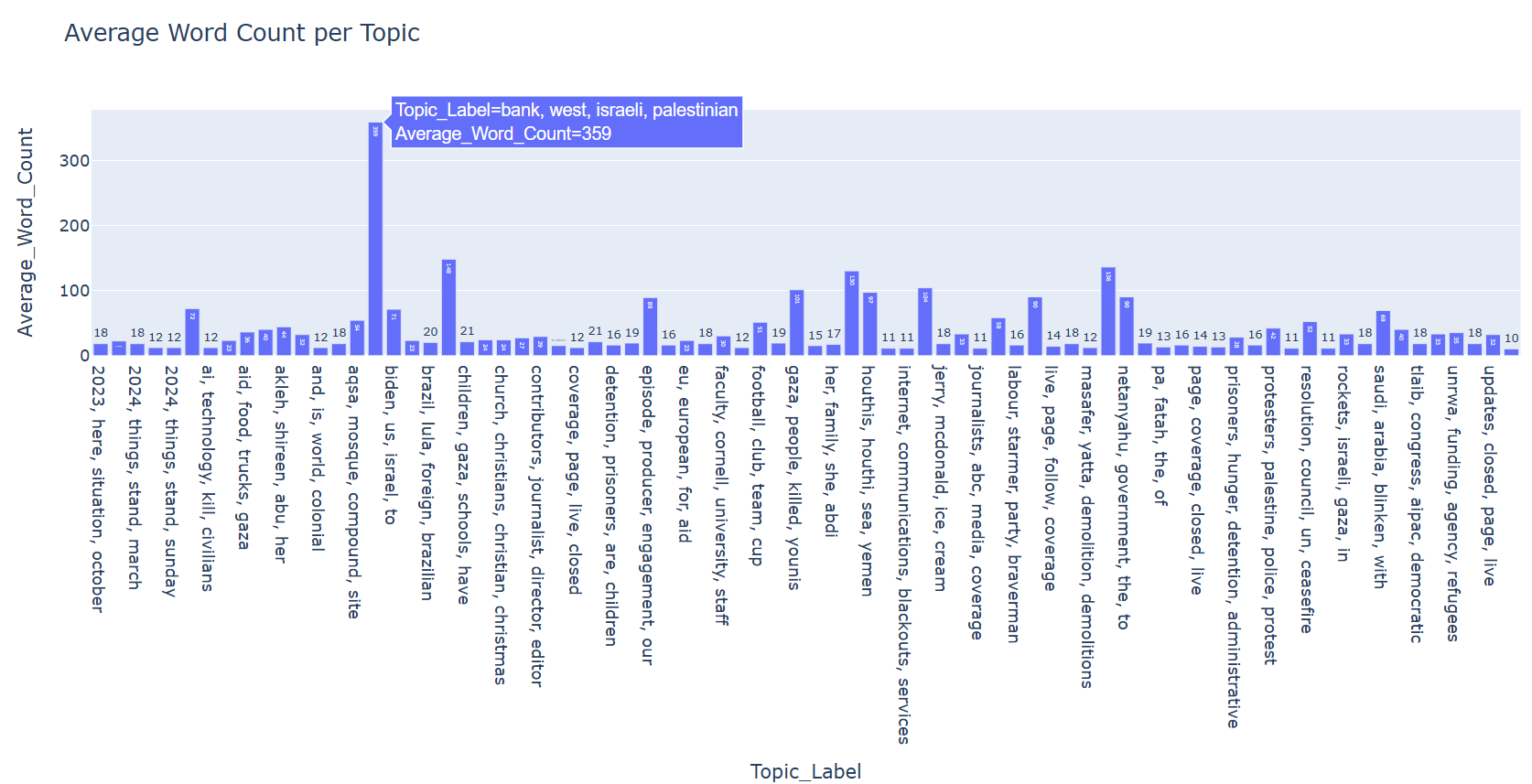
The monthly topic trends graph highlights sharp spikes in article counts which reflect that media coverage is being driven by events and Hamas Isreal war was that event. Form this graph we can observe that after 2023, the topic labeled "captives, Hamas, release, hostages" shows a sudden rise, which highlights they idea that after the uprising of conflict between Hamas and Israel in late 2023 media started to cover this conflict, that is why article counts amplified from late 2023 to early 2024. Similarly, the "hospital, patients, medical, hospitals" topic also increased after October 2023, showing increased media coverage on medical crises, which was because of ongoing war situation. Another important observation made from this graph was the increase in number of articles about "Iran, Iranian, Syria, US", which shows that during this period foreign involvement of regional and international geopolitical system increased, simultaneously increasing global and local media coverage.

**Figure 3: Top 5 keywords per topic**

I also analyzed keywords patterns by plotting another graph which shows the most frequent terms within top 5 topics. This result also reveals that the top articles contain key words like “Israel”, “hostages”, “hospitals”, “Palestinian”, and “Hamas” which reinforces the idea that media articles consistently center around these conflict-driven narratives.

**Figure 4: Average article lengths per topic**

The final visualization I generated was to compare the “Average article lengths per topic”, the result showed that topics about Israeli, Palestinian, and West have highest number of word count and topics including keywords like Hamas, Hostages, Hospitals, Medicals, and Patients are second in order. These results reveal that war-related themes and topics were discussed more in-depth than others, showing how much depth and attention the media gives to these issues.

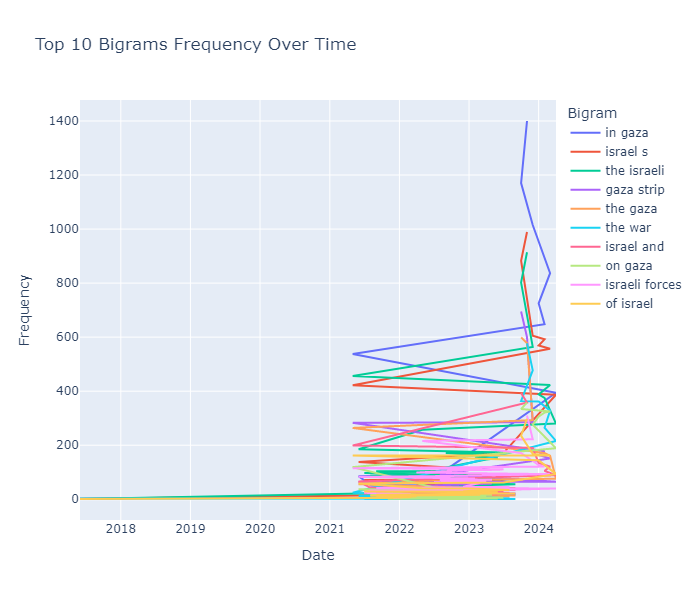


**Main Argument:**By combining topics into themes with LDA and following those themes month to month I could observe that war-related topics for example “Israeli forces / Palestinians”, “captives and hostages” and “hospitals and patients”, certainly dominate the news, and spike sharply after the Hamas-Israel escalation of late 2023. The keyword chart displays this out: the most common words share in the most popular top topics trends are “Israel,” “Hamas,” “hostages” and “hospital,” highlighting how headlines keep returning to the ongoing conflict and its humanitarian fallout. Even the average-length graph tells the same story, because war focused topics have the longest and most detailed articles. According to my visualizations covering the conflict is concentrated around a few key event-driven media storms that prevent a more effective competitive framing from emerging; my observations tend to support our group’s understanding of the coverage of the Gaza conflict as persistent and event-driven.

**Shahzad Rahim (Analysis):**

The data frame I used was N-grams. N-grams are contiguous sequences of words or symbols extracted from text data, it is commonly used in natural language processing and text analysis. An n-gram of size one is called a unigram, size two a bigram, and size three a trigram. These sequences help us to extract linguistic patterns, context, and recurring phrases in large datasets. For example, analyzing bigrams like "Gaza conflict" or "air strike" provides more meaningful insights than single words alone. N-grams are widely used in machine learning for tasks like autocomplete, sentiment analysis, and translation, as well as in trend detection to observe how language evolves over time. By examining frequent n-grams, researchers can identify dominant themes, detect shifts in discourse, and understand relationships between terms.

In my analysis, bigrams were chosen because they offer a balance between specificity and interpretability. While unigrams can be too vague (e.g., "strike" could refer to labor strikes or military strikes), bigrams provide clearer context (e.g., "air strike" or "ceasefire agreement"). This makes them particularly useful for studying media coverage of conflicts, where phrases like "civilian casualties" or "Hamas attack" carry more meaning than individual words. The visualization I used(multi line time series chart) was helpful to present the data as I had to show frequency of bigrams over a period of time. I also visualized the data in a heat map but that wasn’t suitable for the data set I had. In the heat map, it was really hard to distinguish between variables.

I used a list of important words in my code and added a filter to extract the bi grams which had important words in them. I chose words like “Gaza” , “Israel”, and “war” as important words which gave me a frequency of bi grams containing those words. I visualized this data and looked at the time interval in which their frequency started to ascend. It helped me determine when exactly did the conflict escalated as the frequency of bi grams related to war spiked meaning that the articles were talking about the Israel-Gaza conflict.

**Bonus visualization analysis**:

The bonus visualization chart gives the number of articles counts on daily basis, illustrating, sudden spikes, dormant periods and overall trend. Moreover, it complements our main yearly and monthly graphs by presenting a day-by-day level activity. The spikes show whether they align with any war related titles or article length trend.

**A graph with numbers and lines

AI-generated content may be incorrect.**

**Conclusion**

Our group analyzed the Gaza corpus using four methods: Wajahat examined article lengths and war-related words over time, Suleman applied TF-IDF cosine similarity in Gephi to find article clusters, Shehzad analyzed n-gram frequencies, and Sohrab performed topic modeling with LDA. Visualizations showed a rise in war-related terms and article clusters around major conflict events, especially post-2021, confirming media’s consistent focus on the conflict