Gaza October ‘23 to June ‘24 A Critical Discourse Analysis of a Large Corpus of International News Headlines in English

Overview of Research and Purpose

This study aims to examine how various regions across the English-speaking world depict events involving the Gaza Strip in their news headlines from October 2023 to June 2024. Specifically, the objectives are to:

1. Compare the topics related to Gaza that are covered in different English-speaking regions.
2. Analyze how emotional patterns are reflected in the headlines across these regions.
3. Highlight the words and linguistic voices used to describe the events, institutions, and individuals involved.
4. Identify the actions attributed to major entities.

By integrating methodologies from Critical Discourse Analysis (CDA), Corpus Linguistics (CL), and Natural Language Processing (NLP), this study explores how the media employs linguistic strategies to frame the Gaza conflict. Ultimately, the goal is to provide insights into what is being discussed and how it is being represented.

This document supplements the methodology of using BERTopic to narrow-down all the headlines from the NOW corpus from October 2023 to June 2024 to just those relating to events surrounding the Gaza Strip. Please refer to the entire essay “Gaza October ‘23 to June ‘24 A Critical Discourse Analysis of a Large Corpus of International News Headlines in English” for more context. Specifically, this document adds to the methodology that addresses objective 1 above.

One of my goals was to minimize bias by avoiding the use of queries in the text corpus. Topic modeling, an unsupervised machine learning technique, enables the automatic identification and extraction of hidden topics or themes within large text datasets. To achieve this, I employed this natural language processing (NLP) method to address two objectives: (1) determine which headlines are relevant to Gaza, and (2) identify emerging topics.

I utilized BERTopic, a widely recognized and reliable topic modeling method. BERTopic, built on the BERT model, leverages transformers and c-TF-IDF to create dense, interpretable topic clusters while retaining key keywords (Grootendorst 2022; Lemoine-Rodríguez et al. 2024). Using this approach, I classified a dataset of 2,339,216 headlines from the News on the Web corpus (NOW, Davis 2024) into topic clusters. Due to the computational demands of such a large dataset, I divided it into five proportional batches of approximately 470,000 rows each and applied BERTopic to each batch.

BERTopic generates topic representations using a modified c-TF-IDF approach, which treats all headlines within a cluster as a single phrase. This allows it to calculate word importance within each cluster, highlighting the most representative words for each topic. Each topic is characterized by 10 keywords, starting with the most representative word, with each assigned a c-TF-IDF score. A higher score indicates stronger representation of the word within the topic.

Based on the topic cluster visualization, I selected 32 clusters as the final relevant topics. These clusters were isolated, resulting in a refined dataset of 16,065 headlines. A final round of BERTopic was applied to this subset to statistically model the topics, representing events or situations related to Gaza. This process identified 261 individual topics.

To compare reporting across different regions, I grouped the dataset by world regions within the English-speaking realm. This categorization resulted in six regions:

Africa (South Africa, Nigeria, Ghana, Kenya, Tanzania)

Canada/Jamaica (Canada and Jamaica)

Asia (India, Sri Lanka, Pakistan, Bangladesh, Malaysia, Singapore, Philippines, and Hong Kong)

Ireland/UK (Great Britain and Ireland)

Australia/New Zealand (Australia and New Zealand)

United States

Table 3 (Schneeman 2025: 27) provides the distribution of headlines across these regions. After grouping, I applied BERTopic again to extract region-specific topics based on the frequency and relationships between words in headlines from different countries. The results of this analysis, including the region-specific topics, are detailed in Figures 3–7 in the Appendix (Schneeman 2025: 83—87).

Viewing the GitHub Files

When viewing the interactive visualizations resulting from the topic clustering part of this analysis, please consider this following order:

1. Filtering Topic Clusters 1
2. Filtering Topics Stage 1
3. Filtering Topic Clusters 2
4. Filtering Topics Stage 2
5. Final Full Dataset Topic Clusters
6. Final Full Dataset Topic Bar Chart
7. US Topic Clusters
8. US Topic Bar Chart
9. Africa Topic Clusters
10. Africa Topic Bar Chart
11. Asia Topic Clusters
12. Asia Topic Bar Chart
13. Australia NewZealand Topic Clusters

14. Australia New Zealand Topic Bar Chart
15. CanadaJamaica Topic Clusters
16. CanadaJamaica Topic Bar
17. IrelandUK Topic Clusters
18. IrelandUK Topic Bar Chart

Filtering Topic Clusters 1 is the result of the initial BERTopic analysis on the full set of headlines from NOW. The output resulted in the clusters shown in “Filtering Topic Clusters 1,” which were then read through and those clusters encircled in the screenshot “Filtering Topics Stage 1” were then selected and isolated since they proved to be related to Gaza. Next, this subgroup underwent topic modeling again to further filter-out non-relevant topics, which resulted in the clusters in “Filtering Topic Clusters 2.” Again, just those topics encircled in the screenshot “Filtering Topics Stage 2” were selected. Finally, one more round of topic modeling was conducted, resulting in the different topics relating to the conflict in the Gaza Strip from October ’23 to June ’24. These topics are illustrated in their clusters and by their keywords in “Final Full Dataset Topic Clusters” and “Final Full Dataset Topic Bar Chart.” Now, after establishing the headlines in NOW that are discussing events relating to Gaza, a region-specific analysis underwent, which took the final full dataset, subgrouped it by region, and then conducted BERTopic on each regional subdataset. The results are in the region-specific topic clusters and bars attached on this GitHub.

Spreadsheets of Topic Models

In addition to visualizations, this GitHub page also includes the corresponding CSV files for each topic model. These files contain the topic number ('Topic'), the number of headlines per topic ('Count'), the name as topic number plus keywords ('Name'), the keywords ('Representation'), and the representative headlines ('Representative_Docs'). For additional information to supplement the above visualizations, please refer to these CSV files.