UNIVERSITY OF OULU

Faculty of Information Technology and Electrical Engineering

Big Data Processing and Applications

A Study on GDELT Media Representation of the Palestine-Israel Conflict

David Valdivieso (2503261)

Abu Taher (2410233)

Kaisa Anttila (2502847)

Rabiul Hasan (2410234)

May 5, 2025

1 Project Description

In today's world news and information is generated at a much faster pace than is possible for any one human to consume and understand. Different sources can give radically different views to the events they are describing. Because we are forced to choose which sources we follow and which we ignore, we will inevitably end up in an information bubble where the information we get is biased in one way or another. This can significantly affect our opinions and views on different issues. [1]

In our project we want to look into how the reporting of major world events differs in different media. We chose the ongoing conflict between Israel and Palestine as the example event to study more closely. The conflict is reported all around the world, but as it divides opinions, the way the media approaches it can differ. Media reporting of Israel and Palestine has been studied before and has been found to be biased in many news outlets. [2, 3]

We look into this topic using GDELT database. GDELT stands for "the Global Database of Events, Language, and Tones". It gathers news from all around the world into one database and categorises them in several ways to allow for easier search and analysis. We filter and clean the data to represent the Israel-Palestine conflict and create views to the data to analyse the differences in media reports in different areas of the world. We show our results in several graphs and concrete numbers.

2 Related Work

Several previous studies have utilized the GDELT dataset to examine how the Palestine–Israel conflict is portrayed in global media, highlighting the dataset's versatility in analyzing sentiment, bias, and event coverage.

A recent study by Sánchez et al.[4] used hesitant fuzzy linguistic terms to assess European media sentiment during the 2023 Israel–Gaza war, finding Spain's coverage to be the most negative, while Germany and the UK were more neutral. Similarly, Nguyen et al.[5] applied sentiment analysis techniques to the GDELT news corpus to track shifts in media tone during critical global incidents. Their study introduced a method for aggregating sentiment through hesitant linguistic terms, allowing them to detect nuanced shifts in the collective mood across thousands of news articles. Voukelatou et al.[6] used GDELT event data to estimate a "peace index," suggesting media attention reflects political stability. Shaver et al.[7] provided a critical evaluation of how GDELT and other global conflict event datasets cover extrajudicial violence in Israel and Palestine. They found that lower-profile incidents, such as property destruction or non-lethal clashes, were often underreported compared to curated sources like ACLED. This raised concerns about the limitations of relying solely on automated media-based datasets for conflict analysis.

In addition to academic research, several industry projects and journalistic analyses have contributed valuable insights. Media Cloud [8], a prominent open-source media analysis project, examined news coverage trends during the first month of the 2023 Israel-Hamas conflict. Their findings indicated that right-leaning media outlets in the United States devoted significantly more coverage to the conflict compared to other political groups, with noticeable shifts in language over time, such as the increasing use of the term "genocide." A 2024 report by The Nation [9] showed a sharp rise in global media attention to Gaza and the West Bank after the October 7 escalation, illustrating how sudden events can dramatically shift media agendas. Asserson[10] analyzed BBC's English and Arabic coverage of the conflict using GDELT data combined with GPT-4 sentiment classification. The study found a higher level of sympathy toward Palestinians and more frequent association of Israel with terms like "war crimes," prompting broader discussion on media framing and accountability. Leetaru [11] used GDELT's TV Visual Explorer to assess U.S. television coverage, finding brief surges of attention during Gaza escalations but continued prioritization of Ukraine in mainstream media narratives.

3 Data Description

3.1 Overview

In this part, we will first talk about the steps we followed to choose a specific database from GDELT(v1). We will then comment on the preprocessing steps we followed to clean and filter the data, and we will finalize with the description of the final dataset we will use for the analysis.

3.2 Preliminary research

We did an extensive research on what database we would use from GDELT:

At first we looked at the Human Rights Global Knowledge Graph, but it lacked of critical information such as time events.

GDELT 2.0 was promising as we studied its data structure. It had a great potential but it quickly fell out of scope due to its complexity (multiple different delimiters) and the relative size of the files (as one single day could take several gigabytes of data already).

Our solution to keeping the vision we had for this project at an adapted scope was to use GDELT 1.0 as our main dataset. This dataset had the most important qualities for our project:

- Information from news media around the world, although it only gathers information from certain English language media.
- The data schema is simpler compared to GDELT 2.0
- The data amount was more manageable, as there is only one file created per day.

After preliminary research, we estimated that gathering data that represents up to one year and eight months would result in 4.9 GB of zipped files from the GDELT website, which should be plenty for analyzing temporal variations in the data.

3.3 Data Preprocesing

To get the data from the GDELT project, we first scraped the data information from the GDELT event file website. With the filenames that counted with the date and size in their names, we created a data frame to analyze the size of the files that we could download while deciding what range of time we would like to obtain.

Our procedure to treat and prepare the data consisted of keeping only 19 columns out of the original 58. The columns we dropped included data that was redundant or not relevant to our research question and would not improve our analysis significantly.

To filter only material related to the Palestine-Israel conflict, we created a series of search conditions that went through each column if any of these specific terms was included. These conditions were appended in a series of "OR" conditions. We passed this to a pyspark filter function to obtain our data frame filtered with events related only to the two actors from the conflict.

Finally, we wanted to add information about the Cameo codes to make the analysis easier. These are numbers that are given to events covered by media that express what type of event it is. We built a Cameo dictionary so we could add two extra columns: one referencing the section part of a Cameo code (the first 2 numbers of the code) and another one that holds the description of what kind of event this Cameo code means.

Through an iterative process, we would cycle through the files(download, read, delete), do the preprocessing steps and save them into a parquet file. Then we had another function to merge all the files into a single parquet.

In the end we had merged 156 files into a single parquet folder which contained 3 parquet files. The total size of these files in zipped form was 2.09 GB. We ended up using months from October 2023 to March 2024. The total size of data from this time frame was 7.5 GB.

Once we had our main dataset filtered, we still had some more preprocessing to do. For our analysis we needed to know where the news sources were from. For this we utilized an additional GDELT dataset [12, 13] that included the information on where the news media represented by each domain name was from. We joined this dataset to our primary dataset. There was part of the data that was not possible to track (446,644 rows with 1,357 domain names) where they were from, which made us rescind from them.

3.4 Data structure

The final cleaned dataset contained 3,544,505 rows and 23 columns, including the newly added 'domain_name' and 'news_source_country' columns. We were interested in regional abstraction to group the news media, so we also added a region column related to the news source country.

At this point, we had 24 columns with more than 3.5 million rows of data. We used this cleaned and well-structured data for further analysis, stored in 3 parquet files of only 143 Mb.

We could put a summary of our data, but due to its complexity and richness in categorical values, these basic dataset statistics don't provide much information and deeper dive into the data is needed to understand it.

Table 1: Attributes of our final GDELT Dataset

Column Name	Description	Data type
Actor1Name	Actor 1's full name.	String
Actor1Geo_FullName	First actor's full geographic name.	String
Actor2Name	Secondary actor.	String
Actor2Geo_FullName	Secondary actor's full geographic name.	String
IsRootEvent	0 for the root or the main event. 1 if mentioned again.	String
EventCode	3 digit. Same as CAMEO code.	String
EventBaseCode	2 digit base code.	String
GoldsteinScale	Numerical estimate of an event's impact or societal intensity.	Double
NumSources	Distinct media sources that mentioned the event.	Integer
AvgTone	Sentiment score based on textual tone in the news article.	Double
ActionGeo Type	A numeric code for the geographic granularity.	String
ActionGeo FullName	Full location name.	String
ActionGeo CountryCode	The 2-letter ISO country code for the event location.	String
SOURCEURL	URL of the source article mentioning the event.	String
domain name	main media domain from sourceurl	String
news source country	Country from which the reporting media is from.	String
region	Region from which the reporting media is from.	String

4 Methods and Tools

In this project we used PySpark to analyze the data and matplotlib to visualize the results. Besides these two most important libraries, we also used Pandas, socket, geoip2, pycountry, tldextract, geopanda, prophet, and cartopy extensively. Figure 1 describes our overall workflow.

We chose Pyspark because it provides distributed data processing capabilities and can handle a large volume of data efficiently. It is also widely used in industry. Matplotlib is also widely used as it provides figure-making functionality. Pandas is used to convert the Spark DataFrame into a pandas DataFrame for drawing purposes. Socket, geoip2, urllib, and tldextract were used to extract the domain name from the URL

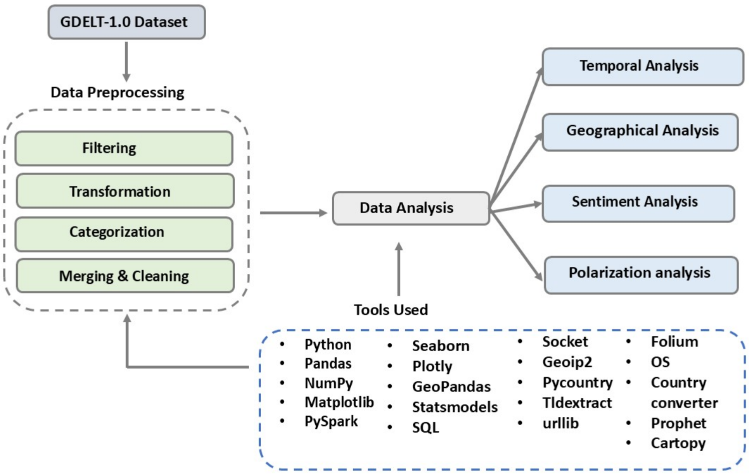

Figure 1: Workflow diagram

and their source country name. Pycountry and cartopy were used for creating global maps. The Prophet model was used to understand the seasonality of news over the period.

To analyze the data, we used exploratory data analysis (EDA) and a machine learning approach. EDA included frequency distribution analysis (e.g., of event types and actors), average sentiment scoring, and temporal aggregation (by month or region). EDA was suitable for this as it us to summarize large volumes of data.

For trend analysis, the Prophet machine learning regression model was implemented. The model helped us to understand the sudden spike and seasonality of the data. The regression-based forecasting approach was chosen for its robustness against missing values and outliers.

5 Data Analysis

In this project, we performed exploratory data analysis (EDA) using various statistics and machine learning methods, such as a Bayesian additive regression model. The analysis was done in cycles. After visualizing and analysing the first sets of data, we used the insights gained from that as a guide for the next analysis.

5.1 Geographical Analysis

Figure 2 depict the number of news items originated form each country. The highest amount of news, around 1,2M, came from the USA. This shows clearly in the figure, where USA is colored yellow for the highest news count. This is explained by the dataset we used as GDELT 1.0 only collects data from English language newspapers. Also the UK and India were major sources of media content; total number of news from these countries was around 0.6M.

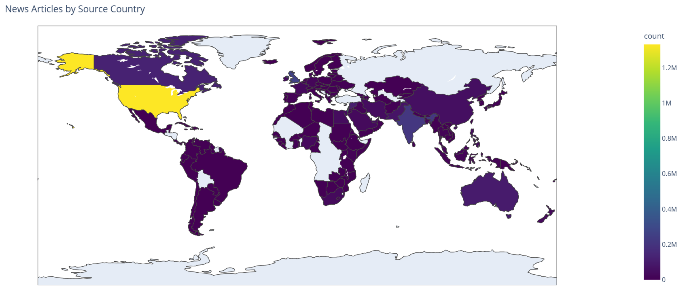

Figure 2: Total news from October 2023 to March 2024 by country

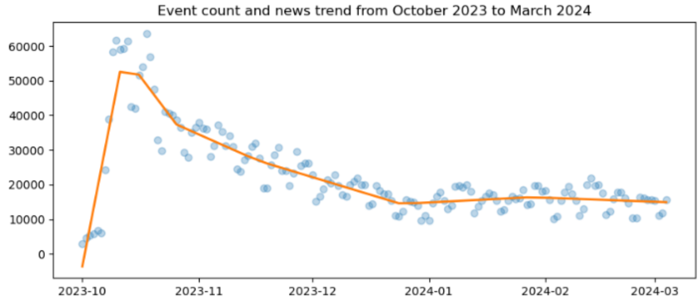

Figure 3: Event count and news trend from October 2023 to March 2024

5.2 Temporal Analysis

We used the Prophet model to understand the news trend. Prophet is a Bayesian additive regression model developed by Meta (Facebook), which takes time series data as input and provides the trend and seasonality of the dataset. We chose this model because it is robust against outliers and has the ability to handle missing values, even though our cleaned data should not have very many of those.

In the figure 3, the blue dots represent individual data points, and the orange line represents the trend. This figure shows the temporal trend of news reporting throughout the period October 2023 to March 2024. From the figure, we can see that reporting spiked on the 7th of October because of the unexpected attack by Hamas, and then it gradually decreased. The lowest amount of news was generated during the period January 2024 to March 2024.

5.3 Analyzing top news sources

We wanted to shed some light on where the data in our dataset comes from to understand it better. The number of datapoints each media in our source data has is not equal to how wide audience that media has. To understand this possible imbalance better we looked at the number of events our data had per each data source and where those medias were located (Figure 4). We compared that list to a list of the most popular and fastest growing domains during the last year (Figure 5) and checked what categories of events those popular domains had reported based on Cameo codes used.

As there is some imbalance in the number of events reported from different regions, we checked the media that has reported the most events:

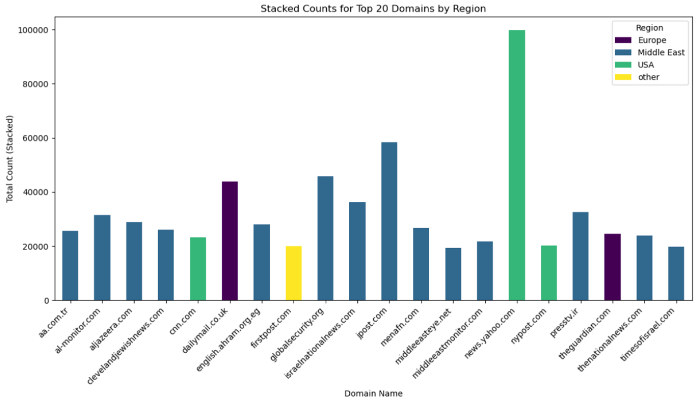

Figure 4: Media Sources with the Most Conflict-Related Reports

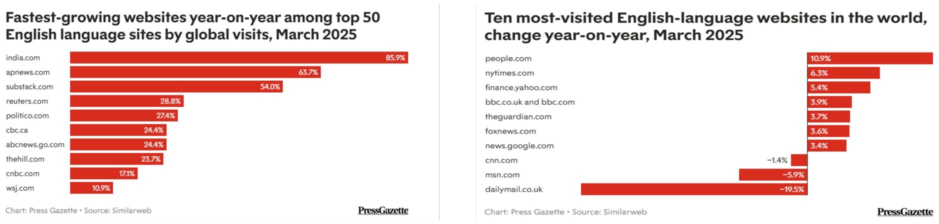
Fastest-growing websites year-on-year among top 50 English language sites by global visits, March 2025

Chart: Press Gazette • Source: Similarweb

Ten most-visited English-language websites in the world, change year-on-year, March 2025

Chart: Press Gazette • Source: Similarweb

Figure 5: Most visited and fastest growing websites in the world

Europe has 2 media in the top 20 and 1128 in total in this dataset. America has has 3 media in the top 20 and 4756 in total in this dataset. Middle-East has has 14 media in the top 20 and 293 in total in this dataset.

As for events America counts with 1,456,141 events, Europe with 426,363 and Middle East with 892,468 (while the region classified as other counts with 892,468).

We compared how the most popular media reported the events (translated in what cameos were most used in the event description) compared with how the events were reported in some of the most active media we have from middle east in our dataset.

While the cameo usage is very similar, we appreciate that there is a slight inclination on more fight codes reported from the middle east media, but not an outstanding difference to clearly state a difference in the report of the events.

5.4 Analyzing top actors

We wanted to take a look at who the most frequent actors are. This also helped to see if our data filter process generated the right data we need for this project. From the figure (Figure 8), we can see that our data filter process was accurate, as most of the actors are related to Israel and Palestine. We further divided the actor pair into two distinct groups, namely state and non-state actors. From the figure, we see that Israel, the USA, Egypt, Iran, and the West Bank are the top 5 state actors. Among non-state actors, the Israelis, Hamas, Gaza, and Palestinians are the most prominent.

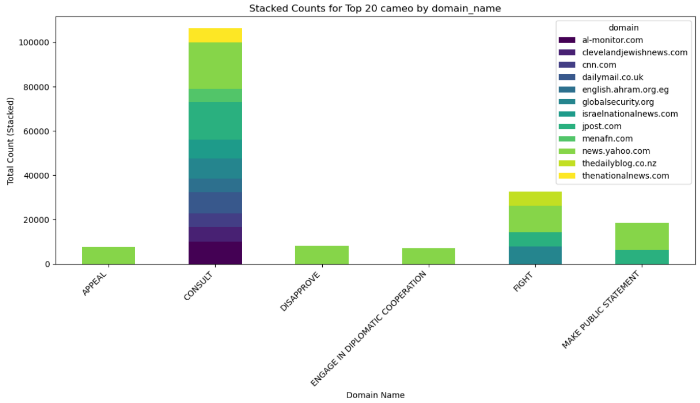

Figure 6: Cameo Categories by Media Source

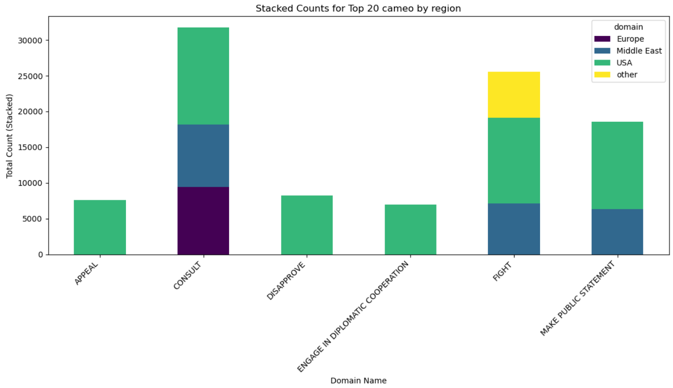

Figure 7: Cameo Categories by Region

5.5 Finding Polarization

To understand polarization, we first wanted to see what kind of news has large differences in average tone in different regions. We did this by looking at one of the attributes our data is categorized by, QuadClass. From Figure 9, it is evident that material conflict is the most polarizing type of news.

Next, we inspected how the average tone for material conflict has evolved during this period in each region. From the figure 10, we see that our initial guess, backed by the data, was right. The red rectangle marks the period during which the most polarization happened. In the Middle East, the average tone was very negative between October 15 and November 15, fluctuating around -5.7, while in Europe the tone increased from -5.4 to -4.8 and USA saw similar change. This coincided with a ceasefire, that apparently was received more positively in the Europe and USA than Middle East. However, the sentiment drops significantly also in Europe and USA as Israel broke the ceasefire in January.

Our next question was, is there a difference in how Israel and Palestine are covered in the news? From the left side of Figure 11, we can see that when it comes to Israel, it was covered very negatively in the Middle Eastern media. Within the period October 15 to November 15, the average tone was fluctuating around -5.9. However, in Europe, it increased from -5.7 to -5.1. In the USA, it was between -5.2 and -5.4.

The right side of Figure 11, also shows a similar kind of trend for coverage related to

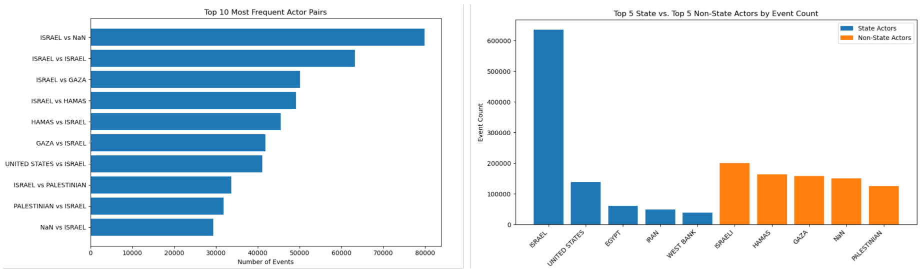

Figure 8: Top Actor Pairs Involving Israel and Palestine

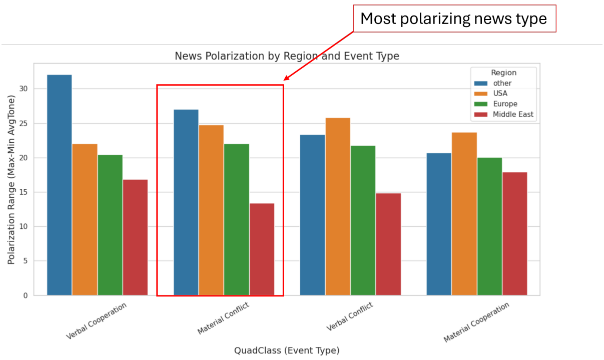

Figure 9: News polarization by region and event type

Palestine, but the difference between Middle Eastern media and other areas is smaller. In the Middle Eastern news, the average tone varied between -5.6 and 6.75. From this data, we inferred that Middle Eastern news has a slightly more positive tone regarding Palestine compared to Israel.

Above all, there is a clear difference in how Israel and Palestine are covered in different regions.

6 Results

Our analysis shows that there are major differences in how media in different areas report the Israel-Palestine conflict. We see that the tone is more negative overall in Middle Eastern media compared to the media in Europe and USA. We can also see how much events change the tone of the news, and that this change is different in different areas.

Our results align with the existing studies. Neureiter (2016) [3] found that media in Britain and Germany held a bias against Israel while the results from the USA were mixed. This showed in our data in how the European sentiment changed more radically over time and in different events compared to the sentiment in the American media. Suwarno and Sahayu (2020) [14] compared coverage of Palestine-Israel conflict in the American The New York Post and the Indonesian The Jakarta Post. Their study saw that both medias framed Israel as the agent of provocateur but The New York Times found Israel's actions more often justified than The Jakarta Post, which has a mostly Muslim reader base. This aligns with our results on the more negative sentiments in the Middle Eastern media, which is also mostly directed at a Muslim audience.

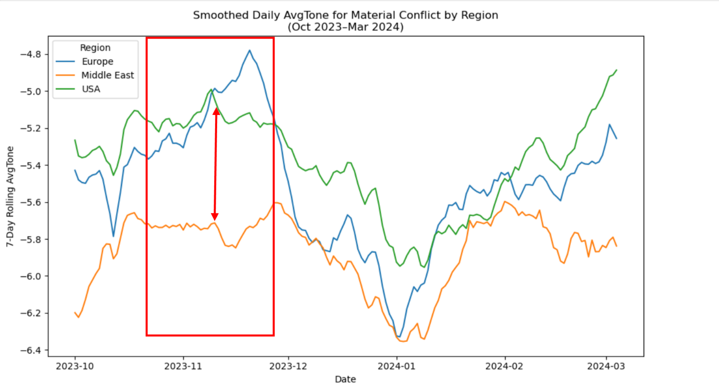

Figure 10: Regional sentiment trend for material conflict.

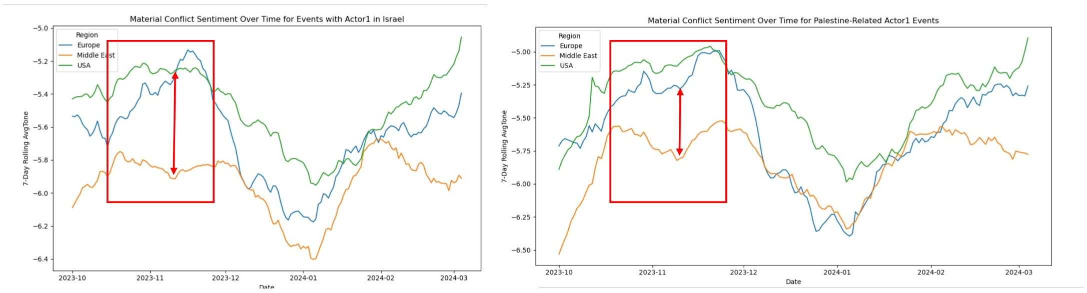

Figure 11: Regional sentiment regarding the coverage of Israel and Palestine.

7 Contribution Report

David Valdivieso: I organized and supervised the project for the most part. I researched about GDELT and its data structure and explored the options for using GDELTV1 and V2. I gathered the GDELT data and did all the preprocessing and filtering (with the exception of adding regions for our domains). I worked on the analysis (related to news media domain temporal and regional distribution) and wrote a part in the final project delivery.

Abu Taher: I have extensively worked on the data analysis part, which includes cleaning the dataset and adding coulmns such as ’domain_name’, ’news_source_country’, and ’region’. I have performed geographical and temporal news trend analysis using a machine learning model, regional sentiment analysis, and polarization analysis. I wrote the data description, data preprocessing, the method and tools sections, and part of the data analysis sections. I also helped write the report draft.

Kaisa Anttila: I helped with planning the project and summarizing the results. I worked on the analysis together with Abu. I wrote the beginning and the end of the project report and made sure the paper was unified in language and style.

Md Rabiul Hasan: I have worked on literature and methodology part. I helped in writing the paper draft.

8 Conclusion

We studied how the media representation of the Palestine-Israel conflict differs in different geographic regions using the GDELT media database. We used a large dataset of about half a year of news between October 2023 and March 2024 that we filtered and cleaned for our use. We then conducted an exploratory data analysis to get insights to several different aspects of the data.

We saw variation in the tone of the news coverage both over time and between different geographical areas. We looked into several different attributes to learn more about our data, and compared our results to existing literature.

This project taught us how to handle a large amount of data. Finding the relevant data among a large database is challenging and requires lots of work and careful consideration. We got experience with using several different tools, such as PySpark, Pandas, and several Python libraries. We handled data in both the original CSV format and in Parquet data storage format. Also data preprocessing became familiar.

We saw some limitations in our analysis. For example, from our data we can clearly see that actors related to Israel are mentioned much more often than other actors. This raises a question about our dataset. Is Israel more active, is its actions more often reported, or does our dataset include news, for example, about the Israel's inner politics in addition to those related to the conflict? As GDELT 1.0 includes mostly English language news sources, we might be missing some essential data from sources that would be in local languages. Thus, these results are an interesting start, but a more comprehensive study could provide useful insights into the subject.

We had some questions we were forced to leave outside the scope of this project. For example, we wondered how reporting might differ inside certain area depending on the media affiliations. Does the right wing media in the US report differently about the conflict than the left wing media? Is there a difference between different European or Middle Eastern countries?

An interesting version of this study would be to do a similar analysis with GDELT 2.0 dataset. Also a wider time range could provide further insight into how the sentiments have possibly changed over time.

References

[1] G. Polyák, Agnes Urban, and Petra Szávai. Information patterns and news bubbles in hungary. *Media and Communication*, 2022. doi: 10.17645/mac.v10i3.5373.

[2] Holly M Jackson. The new york times distorts the palestinian struggle: A case study of anti-palestinian bias in us news coverage of the first and second palestinian intifadas. *Media, War & Conflict*, 17(1):116–135, 2024.

[3] Michael Neureiter. Sources of media bias in coverage of the israeli–palestinian conflict: the 2010 gaza flotilla raid in german, british, and us newspapers. *Israel Affairs*, 23(1):66–86, 2017.

[4] Walaa Abuasaker, Mónica Sánchez, Jennifer Nguyen, Nil Agell, Núria Agell, and Francisco J Ruiz. A comparative analysis of european media coverage of the israel–gaza war using hesitant fuzzy linguistic term sets. *Machine Learning and Knowledge Extraction*, 7(1):8, 2025.

[5] Jennifer Nguyen, Albert Armisen, Núria Agell, and Ángel Saz. Aggregating news reporting sentiment by means of hesitant linguistic terms. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 252–260. Springer, 2020.

[6] Vasiliki Voukelatou, Luca Pappalardo, Ioanna Miliou, Lorenzo Gabrielli, and Fosca Giannotti. Estimating countries' peace index through the lens of the world news as monitored by gdelt. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 216–225. IEEE, 2020.

[7] Andrew Shaver, Hannah Kazis-Taylor, Claudia Loomis, Mia Bartschi, Paul Patterson, Adrian Vera, Kevin Abad, Saher Alqarwani, Clay Bell, Sebastian Bock, et al. Expanding the coverage of conflict event datasets: three proofs of concept. *Civil Wars*, 25(2-3):367–397, 2023.

[8] Ryan McGrady. How the news talked about the israel-hamas conflict in its first month. *Media Cloud*, December 2023. URL `https://www.mediacloud.org/research/how-the-news-talked-about-the-israel-hamas-conflict-in-its-first-month`. Accessed: 2025-04-29.

[9] Alexei Sisulu Abrahams. The missing news about gaza. *The Nation*, February 2024. URL `https://www.thenation.com/article/world/the-missing-news/`. Accessed: 2025-04-29.

[10] Trevor Asserson with RIMe data science. The asserson report: The israel-hamas war and the bbc, 2024. URL `https://asserson.co.uk/wp-content/uploads/2024/09/asserson-report.pdf`. Quoting "The need for impartial and trusted news with no agenda has never been greater." BBC Annual Report 2023/24.

[11] Kalev Leetaru. Television news coverage of ukraine faded a year and a half before gaza took over. URL `https://blog.gdeltproject.org/television-news-coverage-of-ukraine-faded-a-year-and-a-half-before-gaza-took-over/`.

[12] GDELT Project. Gdelt news outlets by country. `http://data.gdeltproject.org/blog/2018-news-outlets-by-country-may2018-update/MASTER-GDELTDOMAINSBYCOUNTRY-MAY2018.TXT`, 2018. Accessed: 2025-05-04.

[13] GDELT Project. Mapping the media: A geographic lookup of gdelt's sources. `https://blog.gdeltproject.org/mapping-the-media-a-geographic-lookup-of-gdelts-sources/`, 2018. Accessed: 2025-05-04.

[14] Suwarno Suwarno and Wening Sahayu. Palestine and israel representation in the national and international news media: A critical discourse study. *Humaniora*, 32 (3):217–225, 2020.