Computational Social Science project

News headlines analysis of the Israel-Palestine conflict

Department of Statistics
Ludwig-Maximilians-Universität München

Marta Caserio

Munich, March 2024

Submitted by Marta Caserio (matr. 12674338)

Abstract

This paper presents a detailed analysis of sentiment, word associations, correlations, and topics discussed in news articles pertaining to the Israel-Palestine conflict. Utilizing datasets categorized as pro-Israel and pro-Palestine, I employ various analytical methods including sentiment analysis, word association, correlation analysis, and topic modeling. My findings reveal significant disparities in language use and coverage between the two datasets, with pro-Israel headlines featuring more negative sentiments, while pro-Palestine headlines exhibit a higher frequency of positive sentiment. Overall, this study sheds light on how media outlets frame and report on the Israel-Palestine conflict, revealing underlying biases, perspectives, and coverage disparities.

Contents

1 Introduction

1.1 Context of the conflict

The history between Israel and Palestine is deeply complex and contentious, marked by centuries of conflict, displacement, and competing claims to the land. It's important to recognize that both Israelis and Palestinians have historical narratives that inform their perspectives on the conflict, and there is no single, universally accepted version of events.

The roots of the Israeli-Palestinian conflict can be traced back to the late 19th and early 20th centuries, with the rise of Jewish nationalism (Zionism) and Arab nationalism coinciding with the decline of the Ottoman Empire. Following World War I, the British Empire assumed control of Palestine under the League of Nations Mandate, with the aim of establishing a national home for the Jewish people while also respecting the rights of the Arab inhabitants.
Tensions escalated in the interwar period as Jewish immigration to Palestine increased, leading to clashes between Jewish and Arab communities. Following World War II and the Holocaust, international support for the establishment of a Jewish state grew, culminating in the United Nations' partition plan of 1947, which proposed dividing Palestine into separate Jewish and Arab states, with Jerusalem as an international city. While Jewish leaders accepted the plan, Arab leaders rejected it, leading to the 1948 Arab-Israeli War.

Due to international actors' high level of interest, media organizations often face criticism for perceived bias in their coverage favouring one side over the other. Numerous studies have been conducted to assess how the conflict is reported in the media, focusing on inherent biases. Given the close relationship between the US and Israel, such biases should come as no surprise. Therefore, it is reasonable to expect that the global media will often defend Israel instead of giving attention to the repeated breaches of human rights against the Palestinian population. In 2002, the Pew Research Center's news interest data highlighted the Israeli-Palestinian conflict as "one of the most followed global news events not directly touching the USA" in its 16-year history (Elmasry, 2009). This demonstrates the immense attention this conflict draws internationally.

The tension between the two factions has been building up over time coming to a breaking point last 7^{th} October 2023. On this date, the terrorist group Hamas conducted multiple attacks in different locations in Israel. These attacks have been the pretext for Israel to start a war to eradicate the group from the Gaza strip. The conflict is currently still ongoing, marking almost six months of conflict which has had brutal consequences for the civilians living in Gaza.

1.2 Why should we analyse the sentiment of news headlines?

Analyzing the sentiment of news headlines about a conflict serves several purposes. Firstly, it helps understanding the overall tone and emotional context surrounding the conflict. Positive sentiment may indicate progress towards peace, successful humanitarian efforts, or diplomatic breakthroughs. Conversely, negative sentiment could signal escalations in violence, humanitarian crises, or diplomatic setbacks.

Understanding the sentiment of news headlines also provides insight into public perception and attitudes towards the conflict. Positive sentiment may reflect optimism or support for particular actors or initiatives, while negative sentiment may indicate frustration, anger, or despair. This understanding can inform policymakers, diplomats, and advocacy groups about the level of public support or opposition to certain policies or actions. By tracking changes in sentiment, analysts can detect shifts in public opinion, media coverage, or the dynamics of the conflict itself. This can provide early warning signs of potential escalations or opportunities for conflict resolution.
Additionally, sentiment analysis can be used to evaluate media bias. By comparing the sentiment of headlines across different news outlets, one can assess whether certain media sources consistently frame the conflict in a positive or negative light or try to push public perception toward one or the other sides of the conflict.

1.3 Structure of the analysis

My analysis is based on news headlines from six countries: the USA, the UK, Germany, South Africa, Saudi Arabia, and Turkey. These countries were selected to highlight differences in news coverage, with some traditionally favouring Israel (USA, UK, Germany) and others leaning towards Palestine (South Africa, Saudi Arabia, Turkey). I wanted to focus my analysis on the difference between news outlets in countries actively supporting one side or the other.

Initially, I polished the data and then examined word frequency. Following this, I conducted sentiment analysis using three dictionaries to capture ranges of emotions expressed in the headlines. Additionally, I explored correlations between the words used and concluded by identifying potential topics within the dataset. Finally, I have fitted a LDA model to get a better understanding of the different topics discussed in this dataset.

In my analysis, I alternatively chose between using the entire dataset or focusing on subsets based on countries supporting different sides of the conflict. In other words, I started by looking at the entire dataset, without considering individual countries. Then, I specifically examined subsets representing countries favouring Israel and those favouring Palestine. This approach provided a comprehensive understanding of both general trends and specific differences in coverage.

2 Methodology

2.1 The dataset

The data that I have used come from Mediacloud, an open-source content analysis tool designed to map news media coverage of current events. It performs five primary functions: media definition, crawling, text extraction, word vectoring, and analysis. The platform tracks hundreds of newspapers, thousands of websites, and blogs, archiving the information in a searchable format.
On this platform, I accessed news headlines from various news outlets across the mentioned countries. I extracted data about the conflict from 7^{th} of October 2023 to the 8^{th} of March 2024, spanning six months of the war, by selecting the headlines containing either "Israel" or "Palestine". For countries other than the UK and the USA, I utilized existing collections of national headlines provided by websites and filtered them for English content. However, due to the substantial volume of data from the US and the UK, I further filtered the headlines by selecting the top five newspapers in each respective country.

For the UK I selected: the Daily Express, the Daily Mail, the Daily Mirror, The Daily Star and The Daily Telegraph. Whereas for the USA I have chosen: The Wall Street Journal, The New York Times, The Washington Post, The La Times and The Boston Globe

2.1.1 Cleaning of the data

As for the preparation of the data, all the versions of the dataset whether they are divided by countries or sides have the same structure: They are composed of around 32,000 units and 9 variables, including the different URLs relative to the article, the language, the source, the date of publishing and the headline text itself.
Once we have our data, it starts the data cleaning; any punctuation marks are removed and then the text is split into individual words, a process called tokenization. This helps later when its needed to see which words are the most frequent.
After that, I get rid of any numbers that might have slipped in, and I also threw out common stop words like "is" and "the". I filtered out specific words related to the conflict, like "Israel" and "Palestine", given their obvious high frequency as the subjects of the analysis.
Finally, it is possible to count how often each word appears in our cleaned-up text. This gives a good idea of the main topics and themes in the news headlines we're looking at. To see even better the frequency of the different words in the dataset we can refer to Figure 1.

Observing this image, it's pretty obvious that it's filled with words about war, humanitarian aid, and the names of government officials from various countries. It's interesting to see "Ukraine" and "Russia" pop up a lot in the word cloud, suggesting there might be some correlatio between different conflicts in the news headlines. After this preliminary analysis, we can go now to the more interesting sentiment analysis.

Figure 1: Wordcloud referring to the complete dataset

2.2 Sentiment Analysis

When people read text, they rely on the emotional tone of words to decide if it's positive, negative, or expresses a more complex feeling like surprise or disgust. We can use text mining tools to analyse this emotional content automatically. One common way to do this is by breaking down the text into individual words (for example using tokenisation or a corpus) and then adding up the sentiment of each word to determine the overall sentiment of the text. It's not the only way to analyze sentiment, but it's a popular method that works well with the tools available.
To measure the actual sentiments of words there are many dictionaries, or lexicons, available which give different categories of sentiments and might go into more or less detail about them. The three most common and the ones that I used in my analysis are:

- NRC: The NRC (National Research Council) Sentiment Lexicon is a lexicon developed by the National Research Council of Canada. It's a list of words categorized by their sentiment, with each word tagged with the emotions it can evoke, such as positive, negative, anger, fear, and so on.. The NRC lexicon covers a wide range of sentiments, including positive, negative, anger, fear, joy, sadness, surprise, and trust.

- AFINN: Developed by Finn Årup Nielsen, a Danish researcher, the AFINN Lexicon is a comprehensive list of English words to which each of them is assigned a sentiment score ranging from -5 to 5. These scores reflect the degree of positivity or negativity associated with each word.

Each of these lexicons relies on unigrams, which are single words. They comprise numerous English words, each assigned scores for positive or negative sentiment, and sometimes for

emotions such as joy, anger, or sadness. For instance, the NRC lexicon classifies words as either "yes" or "no" across categories like positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. Meanwhile, the AFINN lexicon assigns scores ranging from -5 to 5 to words, with negative scores denoting negative sentiment and positive scores indicating positive sentiment.

2.2.1 Bi-grams and correlation analysis

In the field of text mining, the analysis of bigrams and word correlation plays a crucial role in extracting meaningful insights from textual data. We define bigrams as pairs of adjacent words, which if analyzed capture contextual information, allowing a little more precision than the simple analysis of tokens. By considering word pairs rather than isolated terms, sentiment analysis can better capture the subtle nuances and complexities of language, thereby enhancing the accuracy of sentiment classification
This analysis helps reveal important insights about the structure of the text, common phrases, or recurring themes. Furthermore, exploring word correlations is integral to sentiment analysis as it enables the identification of semantic relationships between words. Words that frequently co-occur or exhibit strong correlations may share similar sentiment connotations.

2.2.2 LDA model

Latent Dirichlet Allocation (LDA) is a common algorithm used for topic modeling. It works based on two main principles:

- Each document is viewed as a blend of different topics, with each topic contributing to the document to varying degrees.

- Each topic is seen as a mixture of words, with certain words being more associated with particular topics.

LDA is a mathematical approach that simultaneously estimates these principles, determining both the mixture of words associated with each topic and the mixture of topics describing each document. Various implementations of LDA exist, each aiming to uncover the underlying themes within text data.

3 Results

3.1 NRC Lexicon

The first lexicon which I used is the NRC. In the usage of this, I wanted to look into the different sentiments that might be detected depending on the support of the country of reference. To do that I grouped all the headlines coming from the UK, the USA and Germany under the "pro-Israel" group and the countries of Turkey, South Africa and Saudi Arabia under the "pro-Palestine" label. by then looking at the top 10 positive and negative words of these two groups we can already see some big differences both in the type of language used to describe the same event and also in the main focus of those news.

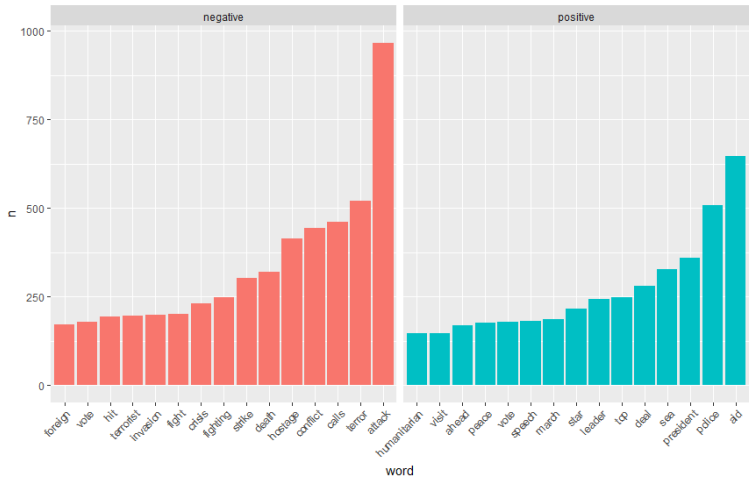

Figure 2: Top 10 positive and Negative words by the pro-Israel dataset

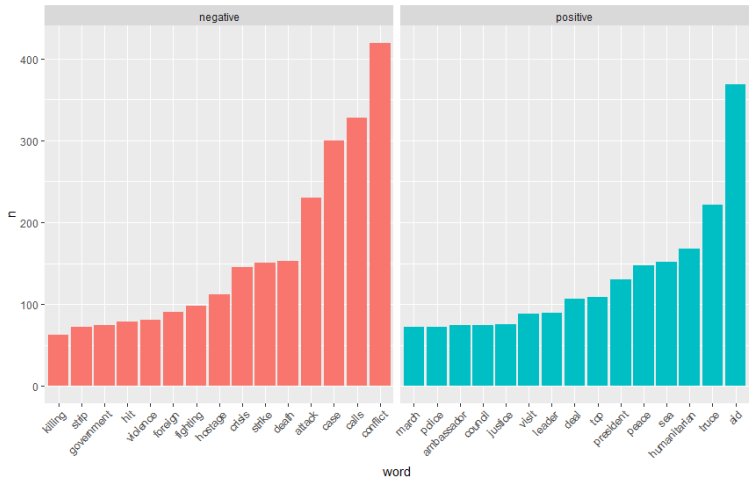

Figure 3: Top 10 positive and Negative words by the pro-Palestine dataset

In both Figure 2 and Figure 3, we're looking at word frequencies sorted into positive and negative categories. What's interesting is the choice of words in the negative category; for instance, in the pro-Palestine dataset, the word "conflict" stands out, while in the pro-Israel dataset, it's "attack." These words mean pretty much the same thing but carry different emotional weights. "Attack" feels more personal and aggressive, while "conflict" sounds more neutral. This difference in word choice hints at the different perspectives of the two groups. One seems to portray events with a sense of direct harm, while the other aims for a more balanced and less biased view.
One thing that we can notice in both of the groups is the presence of the word "aid" as the most frequent word for the positive sentiment.

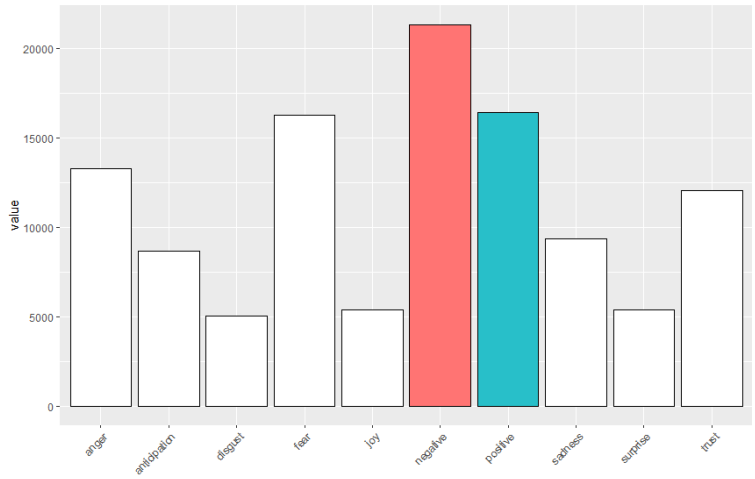

Figure 4: Sentiments detected in the pro-Israel dataset

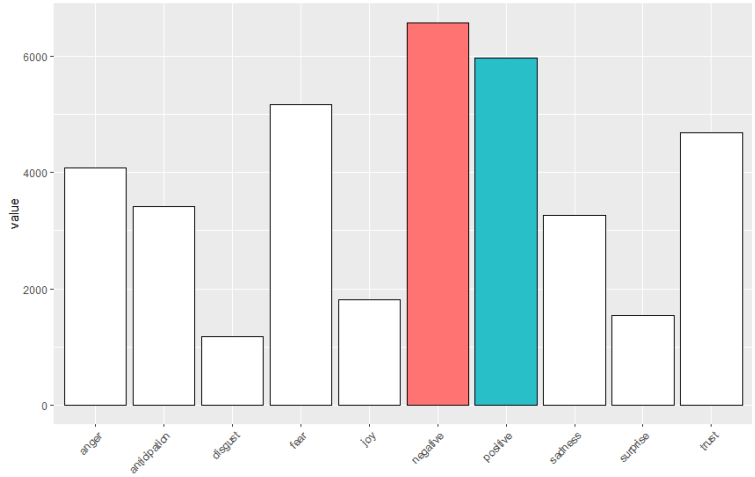

Figure 5: Sentiments detected in the pro-Palestine dataset

Using again the NRC lexicon we can also investigate on the frequencies of the different sentiments which are detected in these groups of headlines. In Figure 4 and Figure 5,

we see the emotions expressed in the headlines, and it's clear that "negative" sentiment dominates, which makes sense given the context of a war conflict. However, when we look closer, there are some interesting differences between the pro-Palestine and pro-Israel datasets. In the pro-Palestine dataset there's more positive emotion with trust as the next emotion, but in the pro-Israel dataset, the emotions classified as positive or fearful are about equal, with anger not far behind.

3.1.1 AFINN and Bigrams

The next lexicon that will be used is the AFINN which, with the usage also of bigrams will allow an analysis more in depth of the words and their associations with each other.

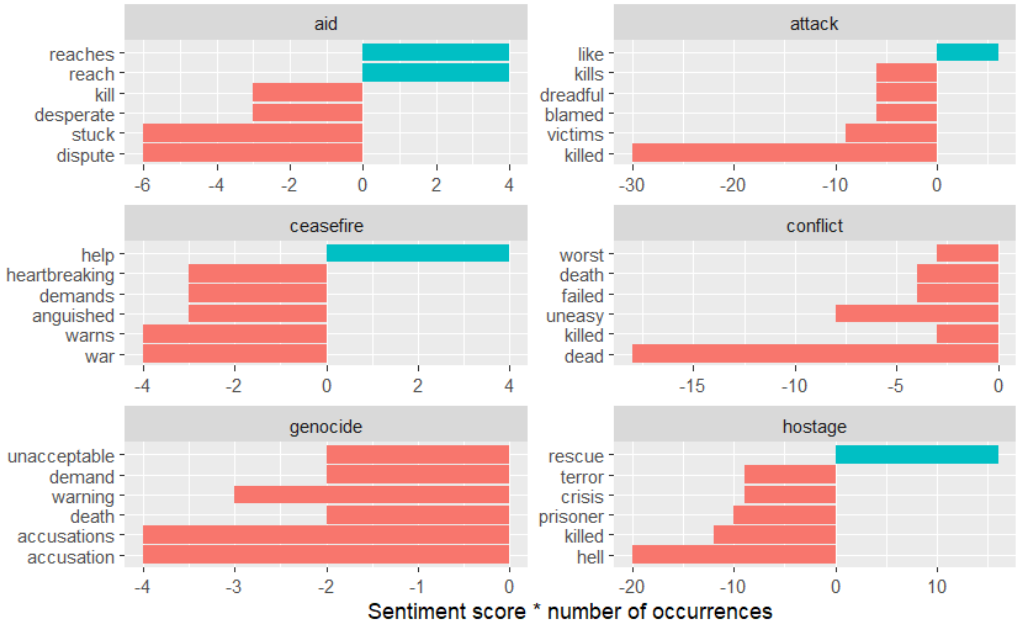

Figure 6: Positive and negative emotion for couples of words in the pro-Israel dataset

In Figure 6 and Figure 7, there are shown the occurrences of different bigrams, or couples, of words with their general sentiment of positivity and negativity. For the two datasets, I have chosen the words "aid", "attack", "ceasefire", "conflict", "genocide" and "hostage". These words are some of the most frequent and might also give some insight into the different perceptions of this conflict. For example we can take the word "ceasefire", which indicates a topic that has been long discussed in international setting but is still refused by Israel. This word in the pro-Israel dataset is associated mostly with negative words such as "war", "warns" and only one positive word which is "help". Conversely, in the pro-Palestinian dataset, the prevalent words include "peace", "justice", "hope", and "agreement", indicating a greater inclination towards ceasefire and a concerted effort towards achieving peace to alleviate the ongoing civilian suffering.

Figure 7: Positive and negative emotion for couples of words in the pro-Palestine dataset

3.1.2 Bigram network

Bigrams can also be analysed without the use of a sentiment lexicon. In this case I wanted to look into a little bit more depth to the connections between the words their most frequent pairings. One of the first things that we can see in these graphs is the

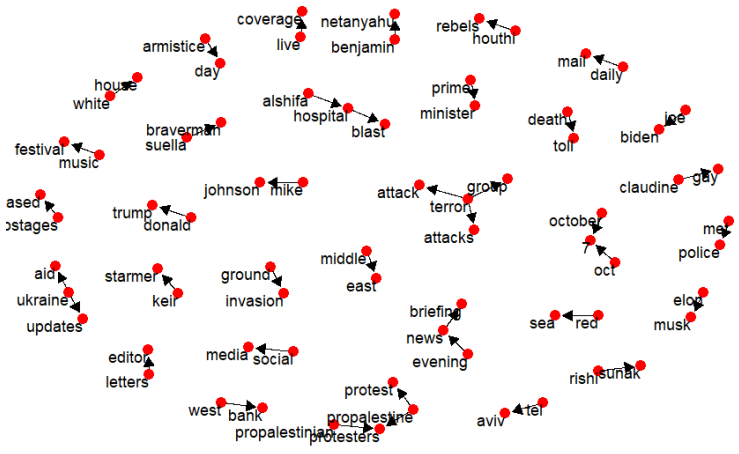

Figure 8: Bigram network for the pro-Israel dataset

huge presence in the Pro-Israel dataset of the names of public figures, these are not only limited to officials of various governments, such as Joe Biden and Rishi Sunak but also popular figures non in government such as Elon Musk or Claudine Gay. The overall high presence of the names of these figures shouldn't surprise at all given the general structure of the media and the highly political and discussed topic we are dealing with.

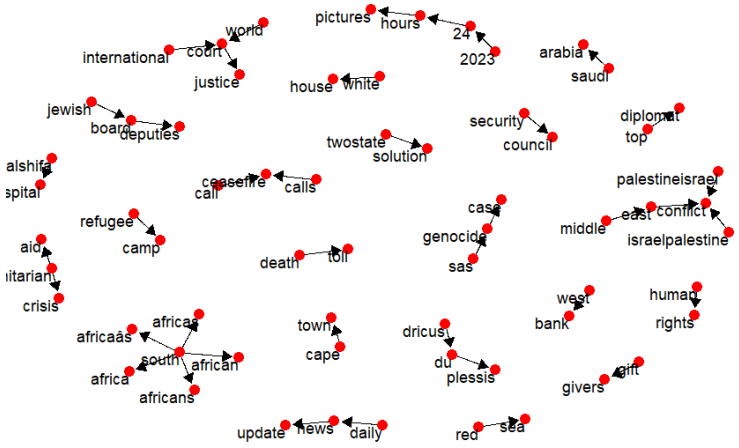

Figure 9: Bigram network for the pro-Palestine dataset

3.1.3 Correlations

Differently to the use of bigrams, when we look at the correlations between words we are not only taking into account the physical proximity of words, but also their frequency in similar topics. This would help us uncovering thematic connections, topical relationships, and underlying structures in the data. In this case I also wanted to look at the different correlations between the same pairings in the two different datasets. I have chosen the words "aid", "attack", "ceasefire", "conflict", "genocide" and "hostage" as you have also seen in section 3.1.1.

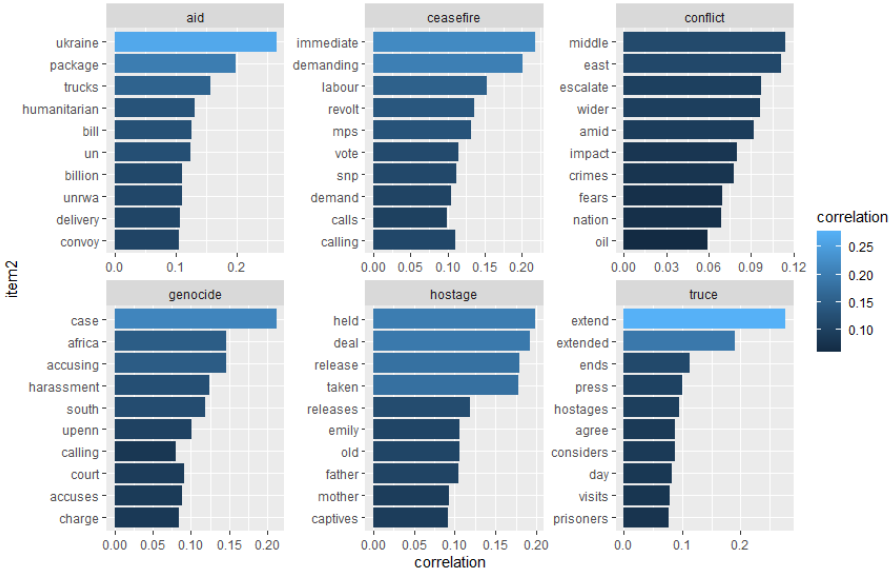

Figure 10: Word correlations in the pro-Israel dataset

Overall we can see that for these words we have higher correlations in the Pro-Israel dataset. The highest correlations are registered for the couples "aid" and "ukraine" and for "truce" and "extend". These refer to two very different topics, the first one is referring

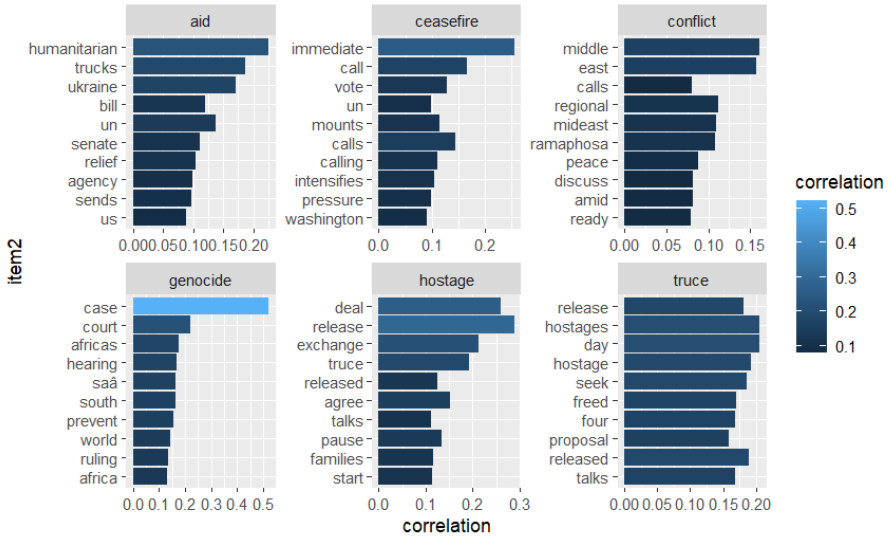

Figure 11: Word correlations in the pro-Palestine dataset

to the continuous parallel between the Ukraine-Russian conflict and the Israel-Palestine one, the second one refers to the small period of truce that was agreed upon between Irael and Palestine during the week between the 24th and the 30th of November 2023. During this period a temporary ceasefire was enforced to allow for the exchange of hostages and allowing the passage of aid toward the area of Gaza.
Using correlations also allows us to create a similar network to the bigram; of course, instead of basing the connections on the frequencies, the links are based on a level of correlation equal to 0.50 (strongly correlated. Looking now at the correlation network,

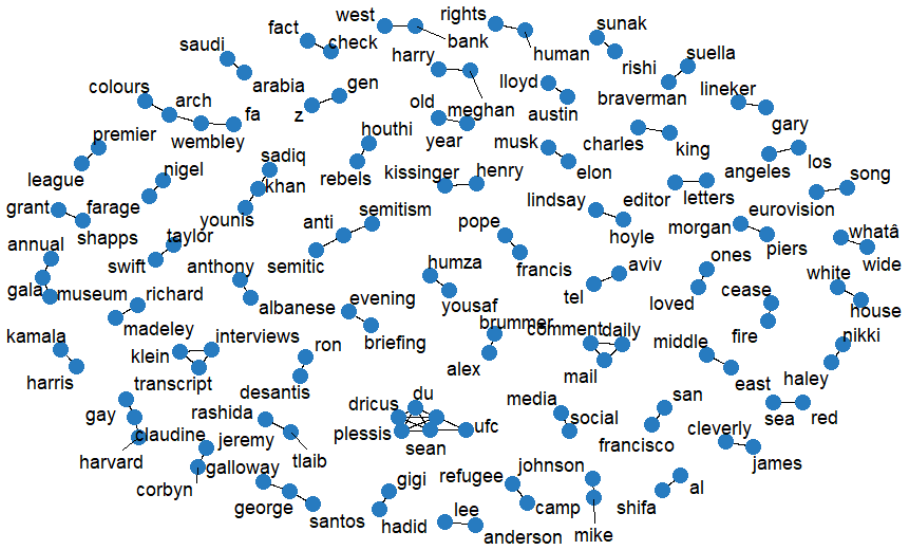

Figure 12: Correlation network in the pro-Israel dataset

the first thing that can be noticed is the different density of words between the two graphs. For the Pro-palestine dataset we have an overall lower correlation between words

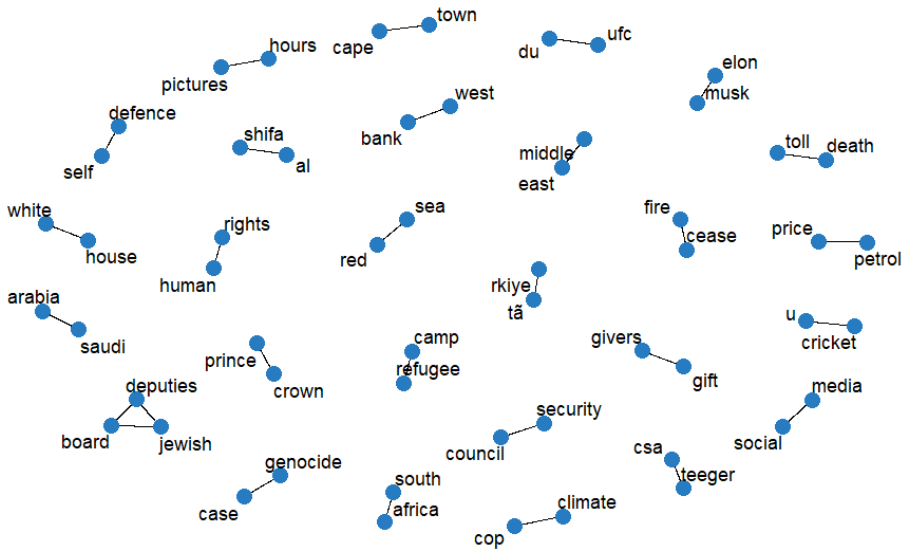

Figure 13: Correlation network in the pro-Palestine dataset

and therefore there are fewer links which have a correlation higher than 0.5. It is also shown that the majority of the words contained in this graph are either words regarding the conflict, such as "humanitarian aid" and "refugee camp", or the names of geographical places, such as "west bank", "red sea" or "middle east".
Whereas for the Pro-Israel dataset the majority of the connections that we can find are names of people belonging not only in the political environment but also of celebrities and other public figures. This happens of course because when referring to a specific person it needs to be specified name and surname, having them always showing up together makes them have a higher correlation.

3.1.4 LDA

As described in section 2.2.2, the use of an LDA model is very useful in text analyses since it allows to see the main topics discussed in the text. Of course the result is not a precise sentence, but rather a collection of words from which the topics need to be extrapolated. In this case, I fixed three topics for this model and used the general dataset as the basis, which ended up being the right number, allocating around 33% of the tokens to each one of them. as you can see in Figure 14 to 16 These three graphs give us a little insight on the topics at hand. I would categorise the three topics as such:

- General information about attacks, strikes, humanitarian issues, military actions and negotiations that are ongoing in the territory of the conflict
- Discussion about justice and human rights in this conflict
- Discussions likely address ongoing tensions, potential conflicts, diplomatic maneuvers, and the involvement of international actors in the region

The second topic is very interesting since it is something that didn't come out as much in the previous parts of the analysis. It deals with the news regarding the different sanctions

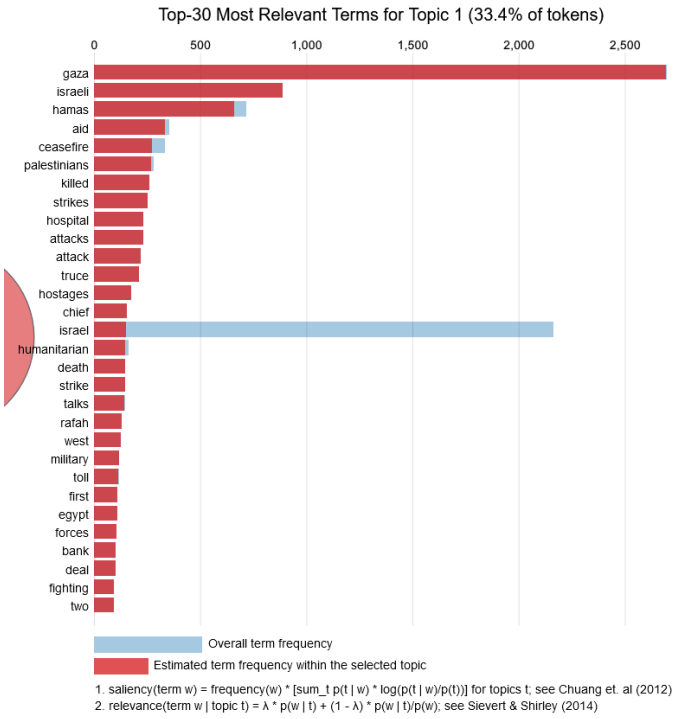

Figure 14: Topic 1

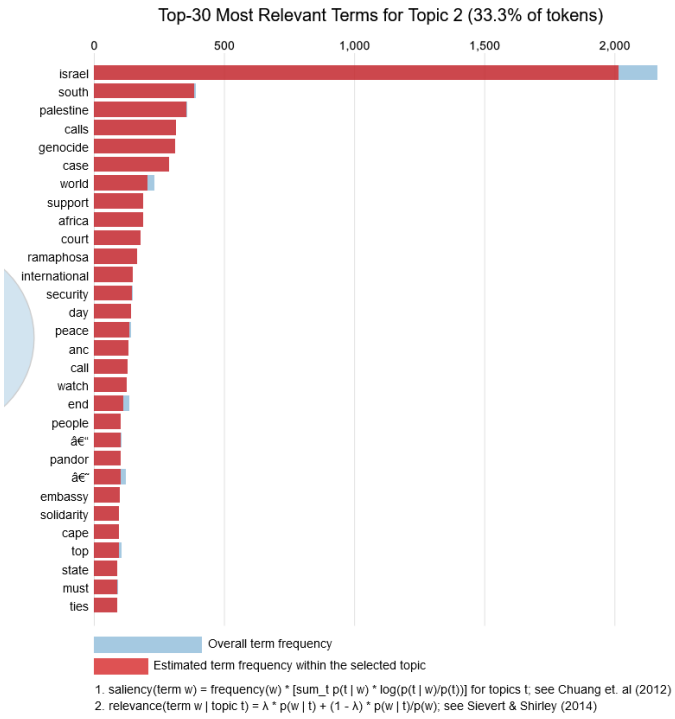

Figure 15: Topic 2

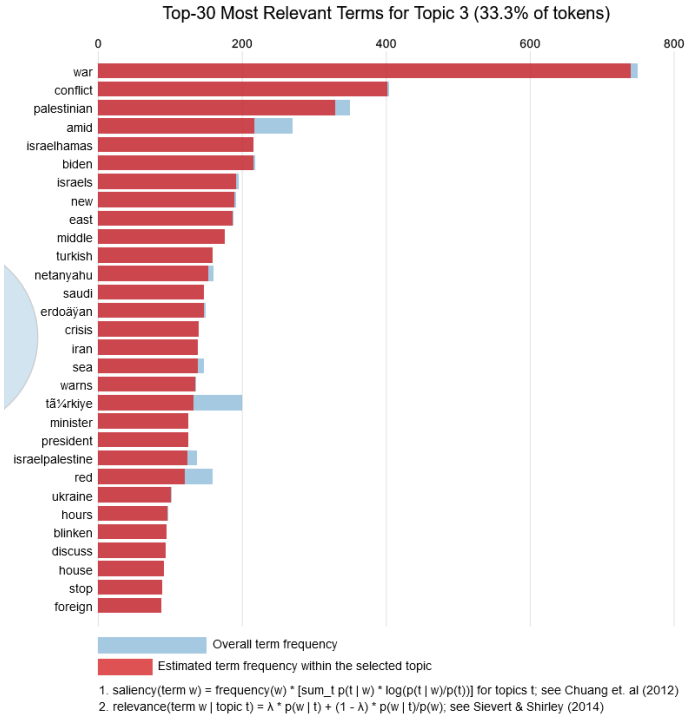

Figure 16: Topic 3

and trials that have been conducted against Israel.
The South African government brought the case against Israel on December 29th 2023, accusing it of "genocidal acts" in its assaults on Gaza. The ICJ (International Court of Justice) ordered Israel to take all possible measures to prevent genocidal acts, to prevent and punish direct and public incitement to genocide, and to take immediate and effective steps to ensure the provision of basic services and humanitarian aid to civilians in Gaza. The court also ordered Israel to preserve evidence of genocide and to submit a report to the ICJ within a month describing out how it is complying with these orders.
As we have see in the other graphs the majority of the news coming from the Pro-Israel dataset tend to not give attentions to this matter and the subsequent human rights violations, focusing more on the other two topics at hand.

4 Conclusion

The text provides a comprehensive analysis of sentiments, word associations, correlations, and topics discussed in news articles related to the Israel-Palestine conflict. It highlights significant disparities between pro-Israel and pro-Palestine datasets in terms of sentiment, with pro-Israel headlines featuring more negative terms like "attack," while pro-Palestine headlines exhibit a higher frequency of positive sentiment, with words like "peace" and "justice" being prominent. This difference in word choice reflects varying emotional weights and suggests divergent perspectives and biases in reporting.

Moreover, the analysis of word associations and bigrams reveals distinct patterns in language use between the two datasets. In the pro-Israel dataset, words like "ceasefire" are associated with negative terms like "war," while in the pro-Palestine dataset, they're linked with positive terms like "peace" and "agreement." These associations shed light on differing views on key topics such as conflict resolution and peace efforts.

Correlation analysis further underscores disparities, showing stronger connections between certain words in the pro-Israel dataset compared to the pro-Palestine dataset. Pro-Israel headlines exhibit higher correlations between words related to ongoing conflicts and geopolitical events, while pro-Palestine headlines focus more on humanitarian aspects and geographical locations. Networks of word correlations highlight the prominence of public figures in the pro-Israel dataset, reflecting the media's emphasis on political leaders and celebrities in reporting.

Topic modeling reveals key themes in the news articles, such as ongoing conflicts, humanitarian issues, diplomatic maneuvers, and international involvement. Specific topics include discussions on attacks, strikes, humanitarian issues, military actions, negotiations, justice, human rights, and diplomatic tensions. The emergence of topics related to justice and human rights violations, particularly regarding sanctions and trials against Israel, suggests a focus on accountability and legal proceedings in the pro-Palestine dataset.

Overall, the analysis provides insights into how different media outlets frame and report on the Israel-Palestine conflict, revealing underlying biases, perspectives, and coverage disparities. Pro-Israel headlines may downplay or ignore discussions on human rights violations and legal actions against Israel, focusing more on military actions and geopolitical dynamics, while pro-Palestine headlines highlight issues of justice, human rights, and international legal actions against Israel, reflecting a focus on accountability and humanitarian concerns.

5 Bibliography

- INTERNATIONAL COURT OF JUSTICE, *Legal Consequences arising from the Policies and Practices of Israel in the Occupied Palestinian Territory, including East Jerusalem*, https://www.icj-cij.org/sites/default/files/case-related/186/186-20240226-pre-01-00-en.pdf

- Suwarno, Wening Sahayu, *Palestine and Israel Representation in the National and International News Media: A Critical Discourse Study*, Humaniora, Vol. 32, No. 3 (October 2020), https://doi.org/10.22146/jh.52911

- Hossain, Karimuzzaman et al.,*Text Mining and Sentiment Analysis of Newspaper Headlines*, Information 2021, 12, 414., https://doi.org/10.3390/info12100414

- Shahzad, F., Qazi, T. A., Shehzad, R. (2023), *Framing of Israel and Palestine Conflict in RT news, Al-Jazeera, CNN and BBC News*, Global Digital and Print Media Review, VI(II), 1-14, https://doi.org/10.31703/gdpmr.2023(VI-II).01

- Elmasry, M. (2009), *Death in the Middle East: An analysis of how the New York Times and Chicago Tribune framed killings in the second Palestinian intifada*, Journal of Middle East Media, 5(1), 1-46.

- A. Macanovic, *Text mining for social science – The state and the future of computational text analysis in sociology*, 2022

- Ted Kwartler, *Text Mining in Practice with R*, WILEY

- https://en.wikipedia.org/wiki/Israeli%E2%80%93Palestinian_conflict

- https://www.cfr.org/global-conflict-tracker/conflict/israeli-palestinian-conflic

- https://www.aljazeera.com/news/2024/2/2/one-week-after-icj-ruling-is-israel-foll

- https://www.ispu.org/wp-content/uploads/2012/02/2012_Palestine-Israel-the-US.pdf

- https://github.com/marcasetta/News-headlines-analysis-for-the-Israel-Palestine-c