Media framing in the news on Israeli-Palestine conflict using NLP techniques

Elif Çanga, Yagmur Helin Aslan

Université Paris 1 Panthéon-Sorbonne

Research Question

The purpose of this analysis is to show how two different countries cover the war in media news and how the reality of war can be reflected in the media in different ways, by using the NLP technique in the context of the Israel-Palestinian conflict. The fact that this issue is linked to international relations and politics has caused most countries in the world to position themselves. In this regard, we conducted a media framing analysis by examining the news of the Qatari media Al-Jazeera and the British media BBC, in which we detected differences in their political statements.

Why We Chose This Question?

As far as the persistent and ongoing conflict between Israel and Palestine since 1948 is concerned, the issue has always got substantial international media attention. However, with the tension escalating following 7 October attacks, an equivalent battle started to be fought over gaining legitimacy. Hence, the language and discourses in the news coverage became more important than ever. A study held by the Glasgow Media Group in 2011 [1], documented the language used by journalists for Israelis and Palestinians reviewing the BBC's news broadcasts. They found that while terms such as "brutal murder", "mass murder", and "massacre" were used to describe the deaths of Israelis, Palestinians were linked with "terrorism". On the other hand, as a global media giant from Qatar, a country with an explicit Pro-Palestinian position with Qatar openly condemning "genocide" in Gaza, Al-Jazeera reported that 2023 was the deadliest year for Palestinians since the 1948 Nakba, while BBC News has recently been criticized for a biased language against Palestine, such as an allegedly deliberate use of the word "dead" for those killed in Gaza whereas the word "killed" for those who lost lives in Israel.[2]

Taking into account all these controversies and the powerful impacts of media framing found in the previous literature, this study aims to portray how the current conflict is conveyed objectively from different perspectives through language usage in media news.

Data and Descriptive Statistics

The news from two different media sources, which are **Al-Jazeera and BBC News**, are collected spanning the time between **7 October 2023 and 31 December 2023**. We cleaned the dataset to make it suitable for NLP packages as you can see from the GitHub link in the footnotes. For cleaning the data, all the non-alphanumeric unique characters including emojis have been extracted and removed from the text. Then we reparsed the sentences after cleaning the texts.

After the data-cleaning process, the total number of news in the merged dataset is **1645** of which **685** news was published on BBC and **960** news was published on Al-Jazeera. Since the BBC website does not allow us to access news links earlier than 12 October, the news from 7 to 11 October has been appended manually using Google search to the previously scraped ones.

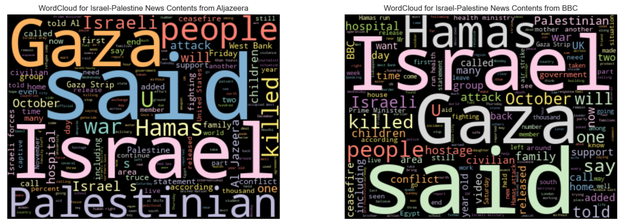

Figure 2. News Contents Wordcloud

Figure 1 reveals a Wordcloud map that visualizes the most frequent word from news headlines in our dataset. As one might anticipate, two most frequent words are "Gaza" and "Israel", followed by words like "Palestinian", "Hamas", "war" and "say". In Figure 2, differently from the headlines, the word "said" seems to acquire a higher relative frequency in the news content, which is quite intuitive. Moreover, the word "Hamas" appears among the top frequent words in BBC news articles while having a lower relative frequency in Al-Jazeera's content.

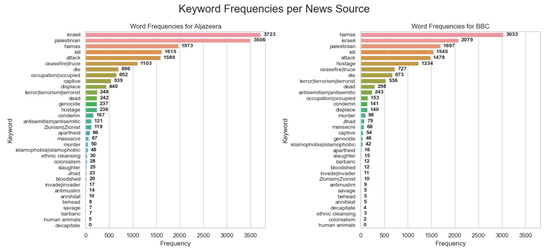

Figure 3. Keyword Occurrences per News Source

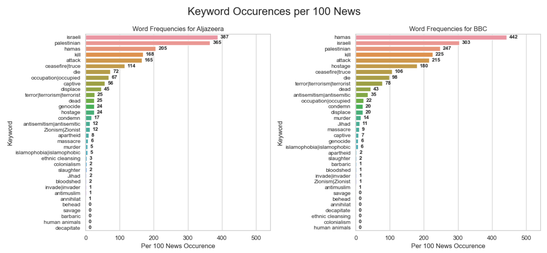

Figure 4. Keyword Occurrences per 100 News

The total number of sentences in our dataset is **40947**. There are **20878** sentences retrieved from BBC and **20069** sentences retrieved from Al-Jazeera. One important indicator of framing and bias in the media is word choices, especially in the coverage of deathly events. In that route, we created a keyword list of words referring to war and death, which frequently appeared in the recent news. Figure 3 which shows keyword frequencies per news source, reaffirms some of our findings from the Wordclouds quantitatively. The figure offers some key insights on word selections: the terms more affiliated with a pro-Palestinian stance such as "genocide" or "Zionism" appear relatively more frequently on Al-jazeera than on BBC whereas terms with a rather pro-Israel connotation such as "antisemitism" or "condemn" have a higher relative occurrence in BBC contents than in Al-jazeera. A significant variation exists between the two sources also on the use of the word "hostage". While the hostages are referred to over **1200** times in BBC content, the word receives nearly 6 times less coverage in Al-Jazeera news.

Interestingly, the latter source seems to display a preference towards the word "captive" rather than "hostage" even in the instances of referring to the same event.

Cambridge Dictionary defines these two words as follows:
captive: someone who is kept as a prisoner
hostage: someone who is taken as a prisoner by an enemy in order to force the other people involved to do what the enemy wants

Hence, these word choices seem like an instance of media framing.
Figure 4 is the adjusted version for the number of news to make a more accurate cross-comparison between the two sources as the sample sizes are different. Different than Figure 3, it shows keyword frequencies per 100 news sources.

Models and Our Approaches

This study used many analyses from the same dataset to capture as many differences as possible between the two media news channels based on the language they used in reporting the news of the Israel-Palestinian conflict. To be able to investigate the media bias and framing, a topic of particular interest is how much coverage fatalities/casualties from each party have received in different news sources. Thus, we attempt to detect, quantify, and visualize the differences in the reference to war casualties of each party in BBC and Al-Jazeera news online. We will deploy two different NLP techniques for a sentence-level analysis among the sentences reporting fatalities: The analyses are as follows:

1. **Identification of the party suffering the casualties using the dependency parsing of SpaCy**. → if the sentence is in active or passive voice what is the subject of the sentence among the sentences that contain the words "killed", "murdered" and "slaughtered
2. **Identification of the party suffering the casualties by fine-tuning a pre-trained LLM (BERT)**. → we designated a text classification task on AI-generated artificial data to train a BERT-based model
3. **Percentage of indirect reporting of the Fatalities**. → we considered indirect reporting as the instances where the news website explicitly states where the given information is coming from.

Results / Figures

1)SpaCy Dependency Tags Method

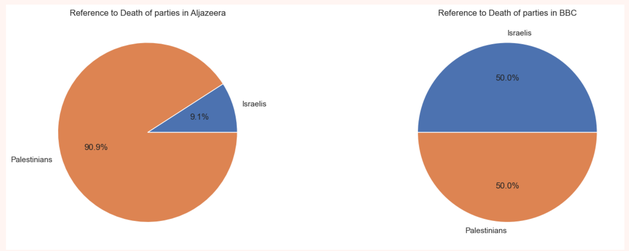

Figure 5. Reference to Death of Parties-SpaCy Dependency Tags Method

In this approach, we use the noun chunks. For each noun chunk in the sentence, we looked for the chunk whose chunk.root.head.text value is one of our three keywords: "killed", "murdered" or "slaughtered". With $chunk.root.dep$ method, we obtained the passive/active voice use of that particular verb. This method normally returns nsubj for active sentences and nsubjpass for passive sentences. Lastly, chunk.text argument in these instances would print out the subject that performs the action. So we searched for the words Israel, Palestine, Israeli, Palestinians, and Hamas within these chunk.texts

Media framing in the news on Israeli-Palestine conflict using NLP techniques

Elif Çanga, Yagmur Helin Aslan

Université Paris 1 Panthéon-Sorbonne

Results / Figures

With this technique, we analyzed **2065** sentences from **970** news. However, this approach managed to identify the party that suffered the reported fatalities only in **500** of the sentences. The reason for this low capacity of identification is due to the fact that SpaCy's noun chunks are because it can only correctly identify the party suffering the fatality when the verb and subject that performs the verb are related to each other by a direct clause.

In **500** sentences, Figure 5 shows that Al-Jazeera refers to Israeli fatalities **9.1%** of the time, whereas BBC reports both sides of fatalities on an equal scale. To the extent these reportings are proportionate to the daily occurrence of attacks and casualties is open to discussion, yet, there seems to be a clear distinction between the two sources.

However, the low number of sentences identified is not the only limit and potential bias of this approach. When using this strategy, we also failed to take into account other conjugations of these verbs such as *("kill, kills, killing, murdering, murders, etc)* as this is a very simple string-based technique.

2)Transfer Learning and Fine-tuning LLM method

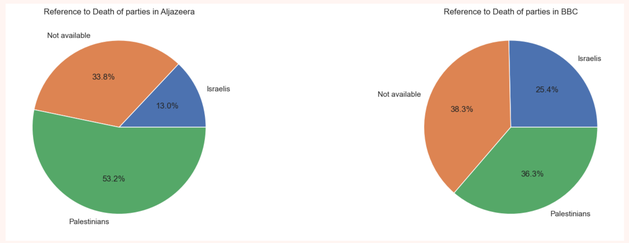

Figure 1. Reference to Death of Parties-Transfer Learning and Fine-tuning LLM method

Known as one of the State-of-the-art (Large Language Models) LLMs, BERT (Bidirectional Encoder Representations from Transformers) revolutionized the NLP by comprehensively understanding context in language through bidirectional training, enabling deeper semantic comprehension with higher flexibility in conjugations. This method can capture relationships between words, encode contextual nuances, and improve performance across various NLP tasks.

First, using several prompts, we asked ChatGPT to generate over 300 sentences referring to the deaths of Palestinians, fatalities from Israeli attacks, and deaths in Gaza. This constituted our training data for sentences mentioning Palestinian fatalities. We augmented our data by replacing the words in the previous sentences with "Israelis", "Hamas attacks", "deaths in Israel" in the appropriate manner to generate also sufficient train data for Israeli fatalities.

Finally, we generated some NA sentences in which the name of the party suffering fatalities is not available. Then, using this artificial dataset with three labels (Israeli fatalities, Palestinian fatalities, Not available), we fine-tuned classic BERT model for our particular task of sentence classification. After a few epochs, the model achieved **0.95** accuracy and **0.95** F1 scores on the validation set.

Our prediction dataset is composed of real-news sentences that have the following keywords: *kill, murder, slaughter, massacre, casualties, fatality, fatalities, die, and dead*. The prediction dataset has **5477** sentences assigned with the labels *'Israelis', 'Palestinians', and 'Not available'*. Analyzing over **5000** sentences, this approach has returned more accurate insight into the news coverages. On the other hand, we still observe a significant difference between the two sources in their coverage of Israeli casualties, which find much less coverage in Al-Jazeera.

Different interpretations may exist on what an unbiased coverage looks like. Is there an equal coverage of Palestinian and Israeli casualties? Does equal coverage imply an unbiased approach? Or should we expect the coverage to be proportionate to the number of fatalities / daily occurrence of fatalities? With several resources, including BBC itself reporting the cumulative death toll of Palestinians to be from 3 to 20 times higher than Israeli fatalities by the end of the year, and with reports of Palestinian casualties each day, we should perhaps interpret Figure 1 in a different light.

3)Direct vs Indirect reporting

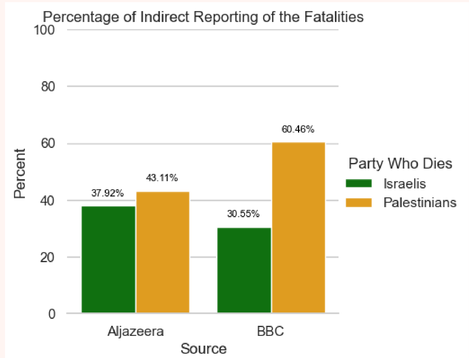

Figure 2. Percentage of Indirect Reporting of the Fatalities

Media framing does not only operate through the number of references/coverages. It is also highly linked to how news is serviced/reported. When skimming the articles on the topic, we identify two distinct ways of reporting fatalities: direct reporting and indirect reporting.

Here, what we try to detect is different from direct vs indirect speech, i.e. whether a third agent is quoted directly or indirectly. What we named direct reporting is the instance of reference to the fatalities without citing any source for that information. On the other hand, we considered indirect reporting as the instances where the news website explicitly states where the given information is coming from. For this purpose, we searched for keywords like "said", "reported", "according to" in the death-reporting sentences.

In the figure above, Al-Jazeera seems to use indirect reporting **37.9%** of the time for Israeli casualties and **43.1%** of the time for Palestinian casualties. The media channel seems to use direct reporting for death tolls, in general, the majority of the time.

On the other hand, when we look at BBC News, we find out that it adopts indirect reporting slightly less frequently than Al-jazeera when it refers to Israeli fatalities, yet it chooses indirect reporting more than half of the time when it comes to Palestinian fatalities.

Since there is a big difference in the reporting style, BBC's uses of direct vs indirect speech seem like less of a reliable reporting strategy than a deliberate media framing. In fact, to get a clearer insight, we actually checked for the sources cited when they used indirect reporting for the Palestinian death toll. We found out that out of **994** sentences in which BBC reported Palestinian casualties, **235** underlined that the numbers were coming from "Hamas-run health ministry". This means that BBC deploys an indirect reporting technique for Palestinian death counts twice as much as for Israeli deaths and almost half of the time, this technique seems to serve to undermine, not increase the credibility of the reporting.

Conclusion

We have scraped **1645** online news on the escalated tension in the Israel-Palestine war from 7 October to 31 December. We have looked for patterns and meaningful analyses on media framing over **40,000** sentences. We find some non-negligible patterns and correlations between the news channel and the reporting language selected.

In particular, we found:
1. BBC News tends to give a higher emphasis on the role of Hamas in the conflict while issuing lower coverage to the Palestinians compared to Al-Jazeera.
2. There are some significant variations in the word choices, like Al-Jazeera's preference for the word "captive" over the word "hostage" or BBC's refraining from the word Zionism.
3. Trained on an AI-generated sentence-level dataset, our BERT-based model predicted BBC to have twice as much coverage of Israeli casualties compared to Aljazeera, while Aljazeera seems to have granted 17% more coverage to Palestinian losses than BBC.
4. BBC used indirect reporting twice as more frequently in reference to Palestinian deaths than Israeli deaths, most of the time in a way to discredit the reliability of these reports with the phrase "Hamas-run health ministry" while Israeli fatalities were overwhelmingly directly reported.

Finally, despite the advances in the usage of NLP in social sciences in general, and in the detection of media bias and framing in particular, these phenomena are still best detected and identified by human readers. Therefore, as a bonus, we designed a special function that prints out random news from the selected media channel of your choice (BBC or Al-Jazeera) and then searches and returns the coverage of the same news in the other media channel whenever available. While you can find the entire code in the GitHub repository, here is an example output (we cut the full content to fit here):

Bonus

The news source from which you want to print a random news: Aljazeera
Title of the news you chose:
UN Security Council passes resolution on increased Gaza aid delivery
Content of the news you chose:
The United Nations Security Council has passed resolution to boost humanitarian aid to Gaza, following several delays over the last week as the United States lobbied to weaken the language regarding calls for ceasefire. The resolution, which calls for steps "to create the conditions for sustainable cessation of hostilities", passed on Friday with 13 votes in favour, none against, and the US and Russia abstaining.The vote came amid international calls to bring the months-long conflict to an end, as Israeli forces pummel Gaza with one of the most destructive campaigns in modern history and humanitarian conditions in the besieged strip reach critical levels . More than 90 percent of Gaza's 2.3 million residents have been displaced, and conditions under Israeli siege and bombardment have been described by UN officials as "hell on earth". ...

Same News in BBC coverage
Title: UN Security Council backs ramping up Gaza aid, but no ceasefire
News: The vote followed days of negotiations to avoid veto by the US, permanent UNSC member and key ally of Israel. UN chief António Guterres said the "real problem" to delivering aid was Israel's ongoing offensive. The UN has warned Gaza is at risk of famine if the war continues. Israel has launched massive military campaign aimed at eliminating Hamas, following the 7 October attack in which militants killed at least 1,200 in southern Israel and took more than 240 hostage. At least 20,000 people have been killed in the Gaza Strip since, according to the Hamas-run health ministry. Friday's resolution was introduced by the United Arab Emirates. Minutes before the vote, Russia - one of the five permanent members of the council - introduced an amendment to revert to an earlier draft calling for an immediate ceasefire. It argued the text gave Israel freedom of movement to further clear the Gaza Strip. The Russian amendment was defeated and both Russia and the US went on to abstain, while the other 13 members of the council backed the text that now calls for creating conditions "for sustainable cessation of hostilities". ...