News Sentiment Analysis

Liza Smaliak

Walsh School of Foreign Service
Georgetown University
es1564@georgetown.edu

Abstract

The onset of the digital age has made the global community more interconnected than ever. With increased accessibility to information, it has become even easier to track events occurring in any part of the world as they unfold. Accordingly, the sources of information with the largest platforms and the fastest publishing time are those which maintain the greatest dominance over the development of global narratives. The nature of the news cycle, which rewards timeliness over accuracy, makes the current state of the media industry problematic, as it has a greater tendency to spread misinformation and perpetuate biases that reinforce a particular worldview [**?**]. This paper seeks to explore how this phenomenon varies amongst prominent news outlets in different countries as global events transpire, with a specific focus on articles published at the start of the Israel–Palestine Conflict in October 2023. The work applies sentiment analysis to evaluate headlines and identify potential biases in their phrasing. Additionally, classification models are built in order to test if it is possible to make predictions based on various features of the data set. The overall goal is to determine whether it is possible to identify particular rhetorical inclinations of a given news outlet, as this could provide further insight into the role of the media in shaping perceptions of global events.

1. Introduction

Media bias is used to describe opinions that are consistently expressed or implied by journalists and news organizations in mass media to convey a certain perspective; it applies to the general character of coverage, as opposed to the opinions expressed by a single journalist or article [2]. Whether intentional or not, bias in an outlet's reporting can have serious implications on the worldview of its audience, especially in regards to coverage of objectively divisive or rapidly evolving news stories, as they tend to fuel more emotional responses.

The strength of the emotional tone relayed in a perspective is described as its sentiment, which is typically categorized as positive, negative, or neutral [3]. In data science, sentiment analysis is a natural language processing (NLP) technique used to determine the sentiment of textual data. The work described in this paper leverages sentiment analysis to contextualize and further explore the data set of chosen articles.

More specifically, this analysis is focused on the reporting of the Israel-Palestine Conflict, which escalated in violence on October 7, 2023, and has continued since. The extracted data focuses on the top articles by dominant news outlets from the first week of the conflict. This topic was chosen due to both its relevance and its polarizing nature amongst the international community. Moreover, given that it is the result of several decades worth of ongoing conflict, news of an unprecedented development was immediately the subject of emotionally charged headlines throughout the world. Initial observations of early headlines, particularly by western media which has historically adopted a strong pro-Israeli stance [4], ultimately motivated the topic of this analysis.

2. The GDELT Dataset

The data for this project was extracted from the Global Database of Events, Language, and Tone. Also known as the GDELT project, it is the largest open database of human society ever created, with over a quarter-billion event records covering the entire world since 1979 [5]. The GDELT project has a wide variety of subsets, two of its primary ones being an event database, which updates daily with a new file storing events by the date they were found in the world's news media and a global knowledge graph, cataloging the human factors that correspond to the event stream (ie, number of protestors, number killed, number displaced or sickened).

This work draws its data from the event database, as it contains the most extensive compilation of published articles. It includes a total of 1,093 articles drawn from 13 news outlets representing various countries and points of view.

2.1. Collecting and Selecting the Data

The news outlets from which the articles were retrieved were intentionally diversified so as to ensure an array of perspectives on the Israel-Palestine conflict were properly represented in the analysis. The goal was to obtain the top 100 articles published in the first week of the conflict by the top news sources (variation of left-leaning, right-leaning, and neutral) from: the United States for its global influence; the United Kingdom for its historical role in the region; Israel and Palestine for their direct involvement. With the given limitations, not all sources had 100 articles listed under them, but there was a desire to include them, which is why they were still incorporated into the dataset. Several additional outlets were considered, but ultimately left out due to language barriers and accessibility issues. The article breakdown per outlet is as follows:

- United States: *The Washington Post* - 100, *The New York Times* - 100, *Fox News* - 100, *Reuters* - 100, *Associated Press* - 37
- United Kingdom: *The Telegraph* - 100, *The Guardian* - 100, *BBC* - 91
- Israel: *The Jerusalem Post* - 100, *YnetNews* - 100, *Haaretz* - 34
- Palestine: *Al Jazeera* - 100, *Mondoweiss* - 30

In terms of the retrieval process, the data was extracted from Google's cloud platform, BigQuery. While there were other possible methods of retrieval, including from GDELT's raw data files or its API, which can be accessed through manipulation of a URL link, accessing the database through BigQuery allowed for the most specificity in the desired extraction. This process required the composition of a SQL query, which ended up using approximately 61 dollars of the 300 dollar credits provided by BigQuery's free trial account by the time all of the desired data had been assembled. The SQL query extracted eight features from GDELT's event database: article URL, average tone, global event ID, date of upload, actor1 code, actor2 code, event code, and a popularity metric. This was completed in 13 separate queries – one for each domain name.

Additionally, the date range was adjusted to only retrieve articles from October 7, 2023 to October 14, 2023; the actor1 and actor2 codes were specified as Israel or Palestine; and the article URL was changed for each domain name.

2.2. Data Pre-Processing

The pre-processing for this dataset was a three-phase series, which included data cleaning, feature engineering, and data transformation. Ensuring that the textual data was properly formatted was essential for being able to conduct an efficient and accurate analysis.

Since the dataset was manually extracted with its desired features, the data cleaning phase only required checking for null values and replacing them accordingly.

Once the data was cleaned, feature engineering was conducted. This was the most fundamental step in the project, as it involved using natural language processing to create two new features: a sentiment category and a sentiment score. The sentiment intensity analyzer built into Python's Natural Language Toolkit (NLTK) was used for this step. Functions were created to analyze the input of URLS, parse the HTML content, and return the compounded polarity of each text with its corresponding score, from most negative (-1) to most positive (+1). If the score was greater than or equal to 0.05, it was labeled as positive; if it was less than or equal to -0.05, it was labeled negative; otherwise it was labeled as neutral.

The third phase of pre-processing was data transformation. This process included two separate steps, cleaning the headlines and standardizing the sentiment score and average tone features. Cleaning the headlines involved lemmatization (the grouping of inflected words); removing phrases (the original column with the titles included outlet names, which were unnecessary and created too much noise); removing special characters, numbers, extra white spaces, and stop words; making every word lowercase and tokenizing it (converting them into smaller parts). A second, unrelated data transformation involved standardizing the sentiment score and average tone values so that they could be used for future comparison.

3. Methods Used

3.1. Preliminary Analysis with Visualizations

In order to get a better understanding of the textual content that composed the dataset, several preliminary analyses were conducted through the creation of visualizations. The first of these visualizations was a set of word clouds based on the headlines of each news outlet, which provided a general overview of the most frequently occurring words within them. These word clouds gave valuable insight into the specific words that publications chose to use in order to appeal to their readers and increase engagement with the content of the articles.

The second visualization used was a bar chart displaying the sentiment scores and average tone associated with each news outlet as a whole, in order to discern if any news outlet in particular seemed to report more extremely on the topic than another. Finally, a third bar chart that displayed the overall distribution of sentiment across the articles was used to provide an overview of the general sentiment expressed by the articles.

3.2. Classification Model Building

With the goal of going beyond rudimentary sentiment analysis so as to see if it is possible to identify particular rhetorical inclinations of a given news outlet, two classification models experimenting with different features were constructed.

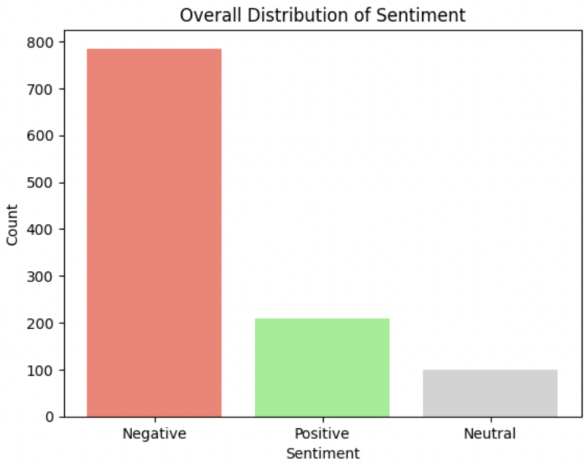


Figure 1: Overall Distribution of Sentiment

The first was a logistic regression model, which was trained on the headlines to predict the sentiment of a news article. In order to perform the regression, a TF-IDF transformation was performed on the headlines. Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure that is used to quantify the relevance of strings within and amongst documents [6], which can be particularly useful in a sentiment analysis. A logistic regression was chosen for its simplicity, interpretability, and efficiency in making predictions from a labeled dataset. The model uses the sigmoid function to map probability values between 0 and 1 according to a given threshold [7].

The second model built was a support vector machine, for the purpose of exploring the prediction potential of features in this specific dataset, as opposed to the more general application that was just described. This model sought to predict an event category based on the sentiment, sentiment score, actor1, and actor2 labels of an article – all of which are categorical variables that were encoded into numerical values for the purpose of the model. Each article in the dataset is attributed to an event code, which represents the action that actor1 performed upon actor2 – these had to be grouped into broader generalizations identified by the GDELT codebook [8]. The SVM model was chosen for its versatility in classification and for its adaptability to smaller datasets. Various kernels were tried, with the final version using a linear one [9].

4. Results

4.1. Visualizations

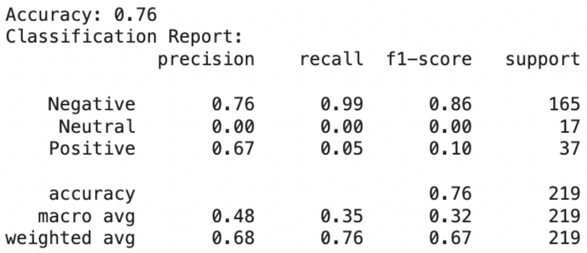
While the visualizations did not illustrate any particularly surprising information, they still provided interesting insights into the character of the data that was collected. For instance, even though the majority of the word clouds displayed a variation of the same words, there was definitely a difference between their prominence, especially amongst the Israeli and Palestinian news outlets. For instance, the most prominent words for Al Jazeera and Mondoweiss were gaza, israel, and palestinian, as opposed to The Jerusalem Post and YnetNews, whose most prominent words were hamas, israel, and terrorist. Word frequency and choice are what shape the portrayal of certain narratives, and this is evident from two opposing perspectives.

The two bar charts did not provide as much internal insight, as the average sentiment score was overwhelmingly negative, which makes sense given the nature of the topic. However, The New York Times did stand out with an average positive sentiment score, inconclusive as to what that may be attributed to.

4.2. Classification Models

The logistic regression model performed with 76% accuracy, indicating a strong ability to predict the sentiment of an article given the weighted relevance of the language used in its headlines. However, it is unclear whether this is transferable to a set of news headlines on completely different topics with more varying sentiments, as this model was trained on a dataset with overwhelmingly negative sentiment scores. While other attempts were made with different models and tuned testing/training sets, most returned errors attributed to the insufficient quantity of data and uneven distribution of labels.

Moreover, the SVM model performed with an accuracy score of just under 18%, proving inadequate in predicting event code based on selected features. This is highly likely to be a result of the quantity of event categories and corresponding lack of supports, which did not provide the model with enough training data to perform at a higher accuracy level. Despite these results, the visualizations that were created to analyze the event code distribution provided some valuable insights, such as the fact that the top three classifications of articles were: “Consult,” “Make Public Statement,” and “Fight,” indicating the nature of the reporting.



```
Accuracy: 0.76
Classification Report:
              precision    recall  f1-score   support

    Negative       0.76      0.99      0.86       165
     Neutral       0.00      0.00      0.00        17
    Positive       0.67      0.05      0.10        37

    accuracy                           0.76       219
   macro avg       0.48      0.35      0.32       219
weighted avg       0.68      0.76      0.67       219
```

Figure 2: Logistic Regression Classification Report

5. Conclusions and Further Work

The visualizations created in this project achieved the goal of demonstrating how the reporting of the same global event varies amongst news outlets, and suggests that there exists a correlation between the source of a story and the manner in which it is reported. While the classification modeling that was conducted was not as successful as anticipated, the shortcomings are clear and can certainly be adapted in further work, as sentiment analysis will become only more relevant as the nature of the news cycle continues to expand.

As mentioned in the results section, the primary shortcomings of the analysis could be attributed to the chosen dataset. One of the initial limitations of collecting this data was that only GDELT's event database was accessible through BigQuery, which restricted the variation of features that could be extracted on an article. Additionally, the decisions that were made in constructing the dataset, which included a limitation on the number of articles extracted from a large variety of news sources, resulted in a dataset that was not as representative or balanced as it could have been. While the reasoning behind these decisions were based on the amount of time it took for the sentiment analysis to process the data, and the desire to analyze a wide range of material, an adaptation of this project could include a focus on a greater quantity of articles from a smaller sampling of news sources.

Furthermore, the subject of the articles that was chosen may have been inadequate for the desired form of analysis, given that the nature of the Israel-Palestine conflict is objectively negative regardless of the perspective that is sharing the story. A limitation of the dataset that could have potentially rectified this is a bias labeling. It was difficult to find methods to automate this process, as machines inherently can not understand the underlying perspectives insinuated by word choice and frequency. This was also a limitation of sentiment analysis, which, while it can provide some insight into the character of a text, can not be depended upon to produce consistent or high accuracy in its labeling.

Further work on this topic is imperative as the nature of international reporting is crucial for the global community's understanding of themselves and of one another. The creation of models that can identify bias or sentiment in a set of text with high accuracy could prove extremely valuable to making not only more informed readers, but global citizens.

6. References