National University
of computer and emerging sciences

Data Visualization and Analysis

Instructor: Ma’am Eisha tur Razia

Project Report

Section: BDS 5B

Roll No	Name
22L-7478	Abdullah Maqsood
22L-7465	Omar Bandial
22L-7541	Eman Atiq
22L-7555	Ameer Tufail

Comparative Linguistic Analysis on Narratives of Mainstream International Media Outlets

Introduction

News articles play a vital role in shaping public opinions and understanding of global events. They are complex and often carry subtle biases, either openly or in hidden ways. These biases and the tone of reporting can influence how events are perceived by readers. Articles use nuanced language, specific word choices, and framing techniques that may shape narratives in ways that are not immediately obvious. Understanding and analyzing these elements require a careful and systematic approach.

This project aimed to analyze the sentiments expressed in news articles from five major global media outlets: BBC, CNN, TRT, Al Jazeera, and Fox News. These sources were chosen to ensure a wide range of perspectives, representing different regions and viewpoints. The project focused on five significant topics: the Israel War, Ukraine War, Islamophobia, US Presidential Elections, and China News. These topics were selected due to their global relevance and the likelihood of generating diverse media coverage.

A dataset of over 5,000 articles was collected by scraping these news sources. The dataset included features such as the article source, headline, description, timestamp, author, region, and full article content. These features provided rich context for each article, allowing for a deeper exploration of how sentiment is conveyed across topics and sources.

The project used advanced data processing and machine learning techniques to classify the sentiment of articles into Positive, Negative, or Neutral categories. Beyond simply classifying sentiment, the project sought to identify key patterns in the language used, uncover keywords influencing sentiment, and explore how sentiments vary across sources and topics. Tools like SHAP were used to make the analysis more transparent and explainable, helping to show why certain articles were classified with specific sentiments.

The significance of this work lies in its ability to uncover the hidden layers of media narratives. Different news outlets may report the same event in ways that reflect their own cultural, political, or ideological perspectives. By combining data analysis with visualizations, this project provides a clear and structured way to understand these dynamics, empowering readers to approach media with a more critical and informed perspective.

Data Collection through Web Scraping

The project began with collecting news articles from five major online media sources: BBC, CNN, TRT, Al Jazeera, and Fox News. The primary objective was to gather a diverse dataset that represented a variety of perspectives and regional viewpoints. Web scraping was conducted using Python and libraries such as **BeautifulSoup** and **Selenium**. The scraping process targeted multiple attributes for each article, including:

- **Source**: The media outlet that published the article.
- **Headline**: The title of the article, often reflecting the main focus or angle of reporting.
- **Description**: A brief summary or introductory text about the article content.
- **Timestamp**: The publication date and time, useful for tracking trends over time.
- **Author**: Where available, the journalist or author of the article.
- **Region**: Geographical information linked to the event or article focus.
- **Full Content**: The complete text of the article, which served as the primary input for sentiment analysis.

Over 5,000 articles were successfully gathered. The diversity of sources ensured a dataset that captured a wide range of reporting styles, tones, and biases.

Data Preparation and Transformation

Raw data scraped from the web often contains inconsistencies and noise. Preparing the data for analysis was a crucial step to ensure accuracy and reliability. This phase involved several cleaning and transformation tasks, including:

- **Removing duplicates**: Articles with identical content or metadata were eliminated to avoid redundancy.
- **Standardizing formats**: Text fields such as dates and headlines were standardized for uniformity.
- **Handling missing values**: Missing data in fields like author or region was either imputed or marked as unavailable.
- **Feature creation**: New columns, such as word counts or sentiment placeholders, were added to support further analysis.

This structured dataset ensured consistency and quality, forming a strong foundation for the subsequent steps.

Sentiment Classification Model

To determine the sentiment of each article, a **transformers-based sentiment analysis pipeline** was employed. The model was pre-trained on a vast corpus of textual data and fine-tuned using Twitter sentiment analysis parameters to better understand shorter and contextually rich text. The sentiment model categorized each article into one of three classes:

- **Positive**: Articles conveying optimism, praise, or favorable tones.
- **Negative**: Articles reflecting criticism, pessimism, or adverse tones.
- **Neutral**: Articles that maintained a balanced or impartial tone.

The sentiment classification added a new column to the dataset labeled "Sentiment." This enabled a detailed exploration of how sentiment was distributed across topics, sources, and time.

Keyword Extraction

Understanding why an article was classified with a specific sentiment required transparency in the model's decisions. For this, **SHAP (SHapley Additive exPlanations)** was utilized. SHAP identified key words or phrases within each article that significantly influenced its sentiment classification. For example:

- In a **Positive** article, words like “success,” “peace,” or “cooperation” might play a dominant role.
- In a **Negative** article, terms such as “violence,” “failure,” or “controversy” might be influential.
- **Neutral** articles often contained descriptive or factual terms that lacked emotional weight.

This step provided deeper insights into the patterns of language usage that shaped sentiment classification and helped identify potential biases in how topics were discussed.

Machine Learning Model Testing

To validate the sentiment classification process, traditional machine learning models were tested on the same dataset. These included:

- **Naive Bayes**: A probabilistic model well-suited for text classification tasks.
- **Logistic Regression**: A simple yet effective algorithm for binary and multiclass classification problems.
- **Decision Tree**: A model that splits data based on feature importance to make predictions.

Each model was trained using the sentiment labels generated by the transformers pipeline, and their performance was evaluated using metrics such as accuracy, precision, and recall. Comparing these models ensured that the chosen method was both reliable and robust.

Visualizations

To make the analysis understandable and engaging, data visualization was a critical step. Various tools, including **Seaborn**, **Plotly**, **Streamlit**, and **Power BI**, were used to create informative and interactive visualizations:

- **Bar Charts**: Displayed sentiment distribution across topics and media outlets.
- **Pie Charts**: Illustrated the proportions of Positive, Negative, and Neutral sentiments within the dataset.
- **Word Clouds**: Highlighted frequently occurring keywords in articles associated with each sentiment.
- **Line Charts**: Showed trends in sentiment over time, revealing how media tone changed during significant events.
- **Bivariate Dot Charts**: Explored relationships between variables, such as region and sentiment.
- **Map Charts**: Visualized geographic patterns in the dataset, showing where specific topics were more prominently covered.

These visualizations were not only valuable for analysis but also played a key role in presenting findings to stakeholders.

Results and Outcomes

The analysis produced several important findings:

- Sentiments varied significantly by topic and source. For instance, articles on the **US Presidential Elections** had a higher proportion of Negative sentiments, while **China News** showed a more balanced distribution.
- Keyword analysis revealed recurring terms that shaped sentiments, providing transparency into the classification process.
- Media outlets differed in their tone and focus when covering the same topics, reflecting potential biases.
- Sentiment trends over time showed shifts in tone, especially during major developments in global events.

These insights were deployed through an interactive **Streamlit application** and a polished **Power BI dashboard**, making them accessible to a broader audience.

Future Improvements

The project can be further enhanced by expanding the dataset to include additional sources and languages.
Incorporating advanced natural language processing techniques and real-time analysis could also provide
richer insights. Additionally, exploring context analysis and framing in articles could offer a deeper
understanding of media narratives.

Conclusion

This project demonstrated a comprehensive approach to analyzing sentiments in news articles, offering valuable insights into how media outlets frame and communicate major global topics. By collecting over 5,000 articles from diverse sources, the analysis captured variations in reporting styles, tones, and potential biases across outlets like BBC, CNN, TRT, Al Jazeera, and Fox News. The thorough data cleaning and structuring ensured the dataset's quality, forming a solid foundation for analysis.

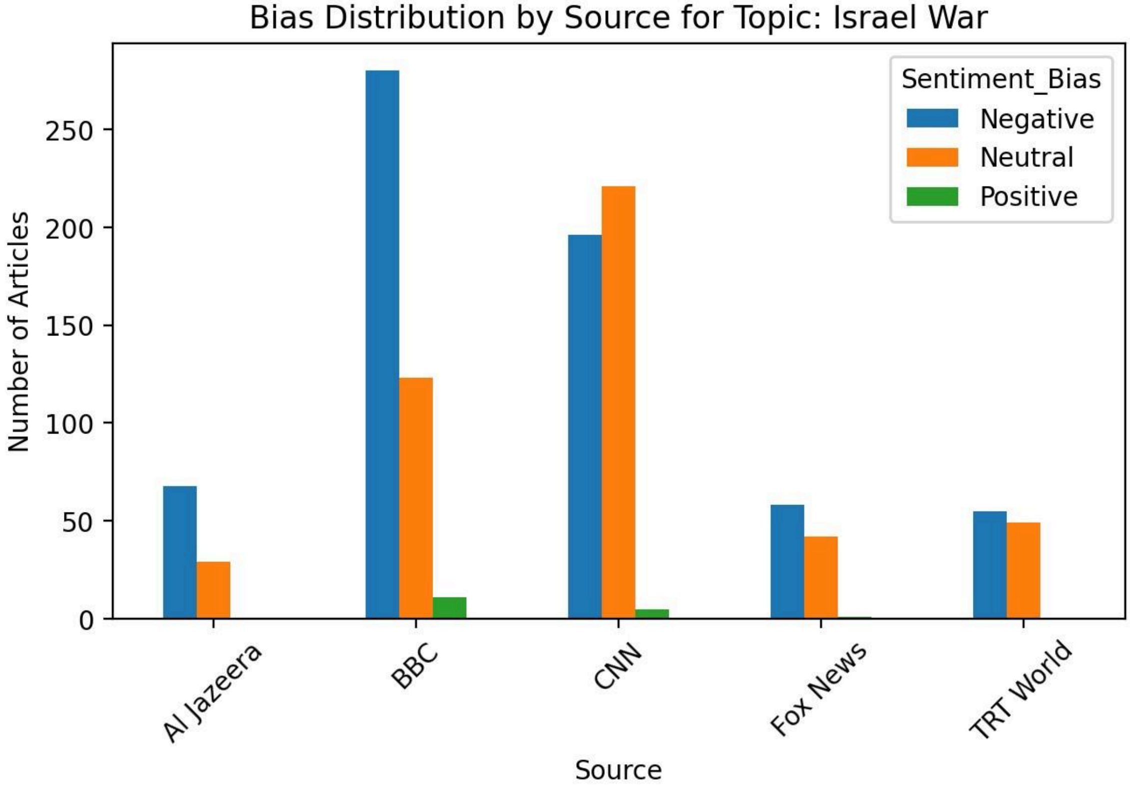
The sentiment classification, powered by a transformers-based model, accurately labeled articles as Positive, Negative, or Neutral. SHAP further enhanced transparency by identifying keywords influencing sentiment, providing a deeper understanding of language patterns. Comparing traditional machine learning models like Naive Bayes and Logistic Regression validated the robustness of the classification approach.

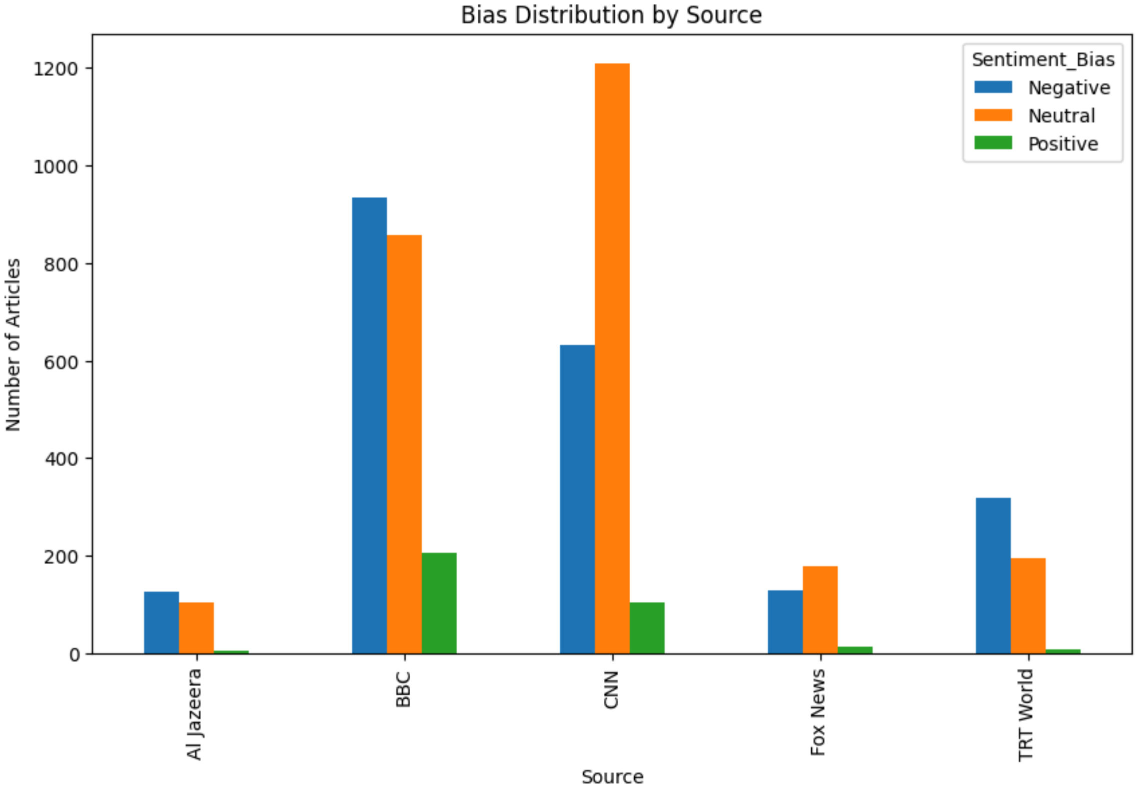
Visualizations created using Seaborn, Plotly, Streamlit, and Power BI brought the findings to life, showcasing sentiment distributions, keyword trends, and regional patterns. Interactive tools made the insights accessible and actionable for deeper exploration. Key findings highlighted significant differences in how media outlets report on topics such as the Israel War and US Presidential Elections, revealing shifts in tone and framing across sources.

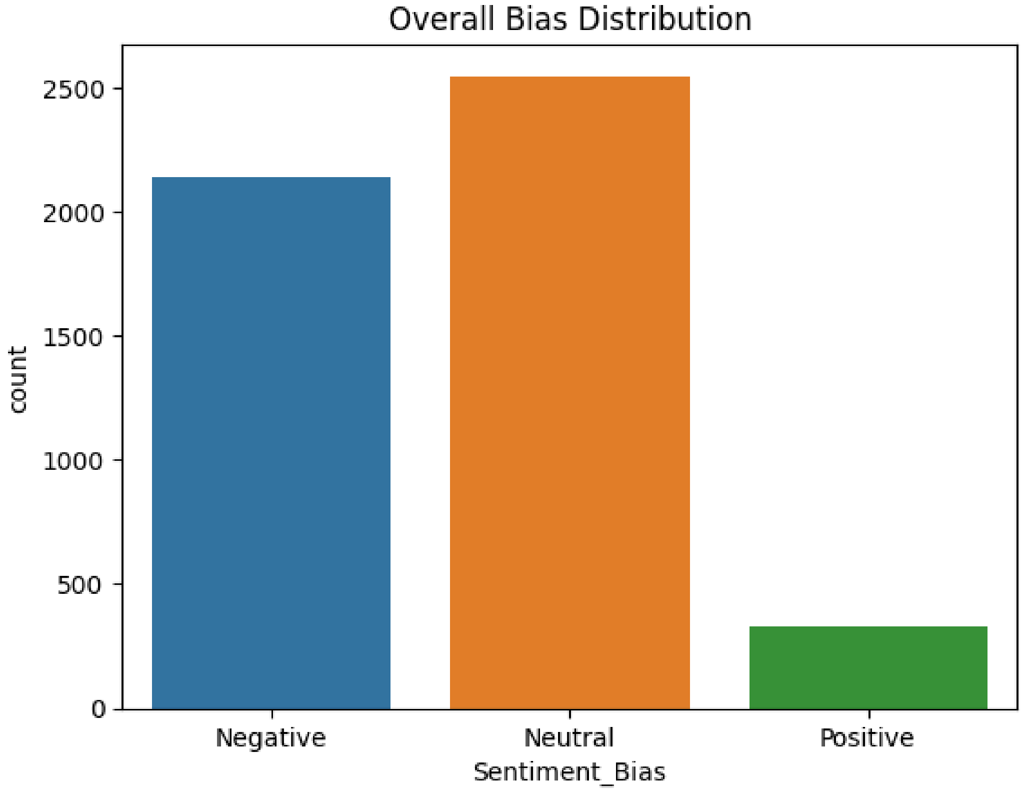
This project underscores the importance of critically engaging with media content to understand biases and nuanced language. Future enhancements could include expanding data coverage, incorporating real-time analysis, and exploring advanced techniques to uncover hidden biases.

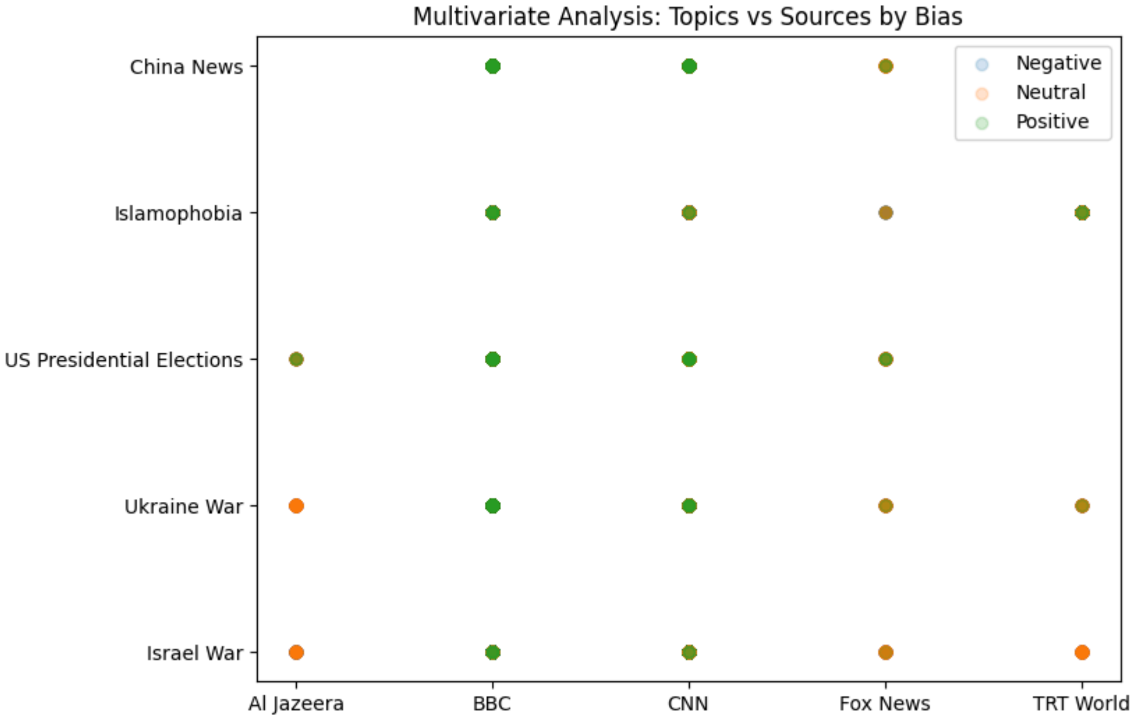
In conclusion, this project successfully provided a framework for sentiment analysis and media narrative exploration, highlighting the role of media in shaping perceptions. With further development, it has the potential to become a powerful tool for monitoring and analyzing media reporting on a global scale.

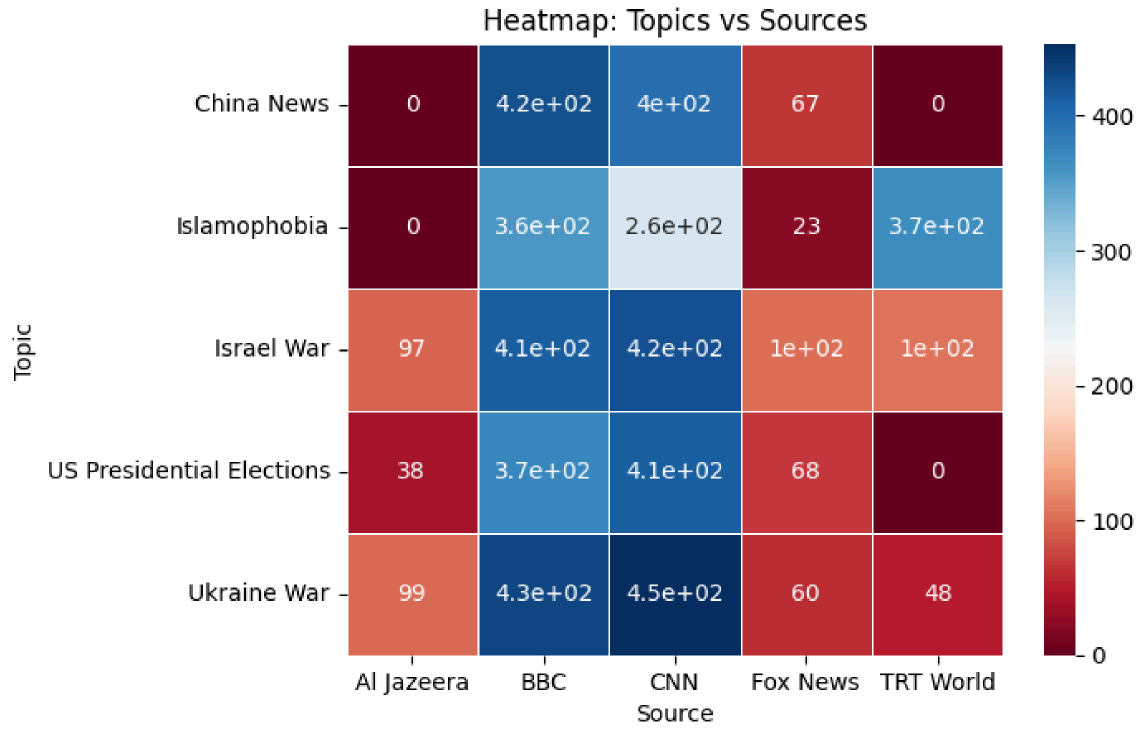
Visualizations

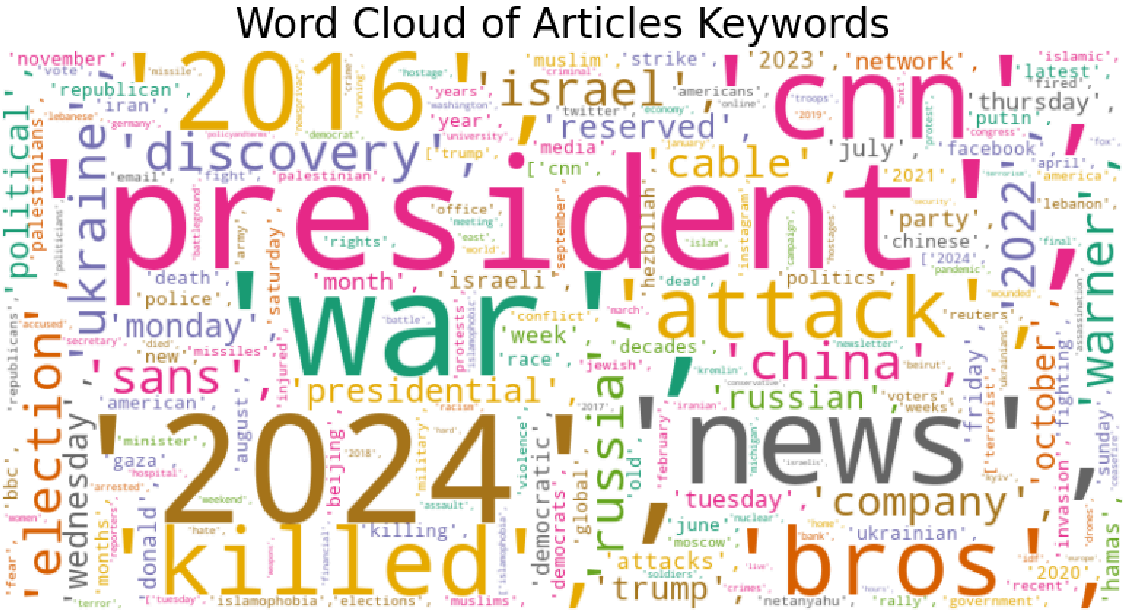




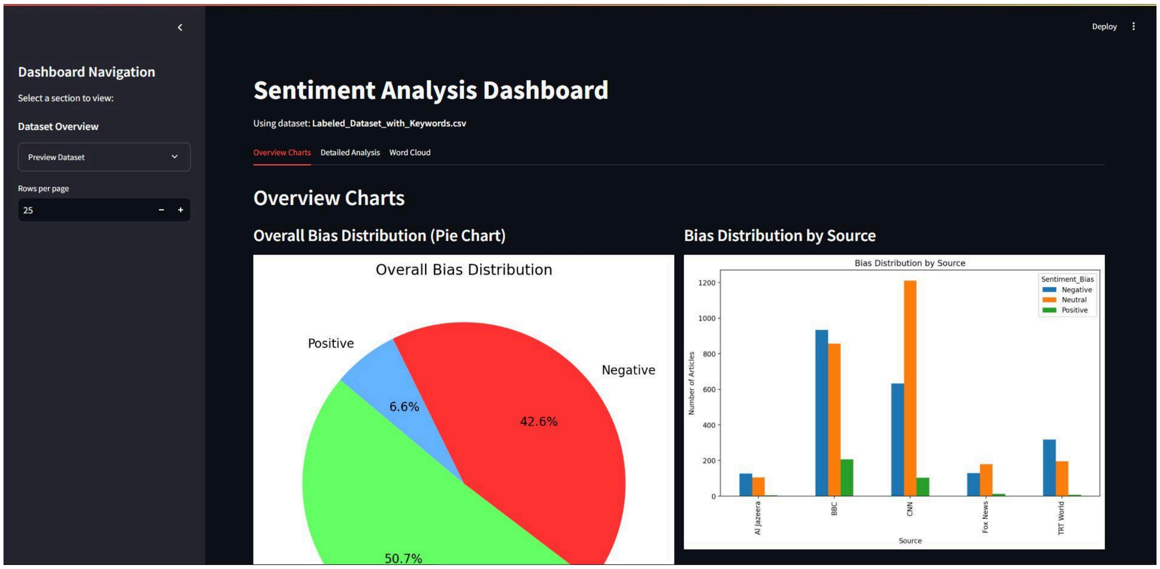

Overall Bias Distribution
count
2500
2000
1500
1000
500
0
Negative
Neutral
Positive
Sentiment_Bias


Multivariate Analysis: Topics vs Sources by Bias
China News
Islamophobia
US Presidential Elections
Ukraine War
Israel War
Al Jazeera
BBC
CNN
Fox News
TRT World
Negative
Neutral
Positive


Heatmap: Topics vs Sources
Topic
China News
Islamophobia
Israel War
US Presidential Elections
Ukraine War
Source
Al Jazeera
BBC
CNN
Fox News
TRT World
0
4.2e+02
4e+02
67
0
0
3.6e+02
2.6e+02
23
3.7e+02
97
4.1e+02
4.2e+02
1e+02
1e+02
38
3.7e+02
4.1e+02
68
0
99
4.3e+02
4.5e+02
60
48
400
300
200
100
0


Word Cloud of Articles Keywords
'president'
'war'
'2024'
'news'
'2016'
'cnn'
'killed'
'bros'
'attack'
'discovery'
'israel'
'ukraine'
'china'
'election'
'russia'
'warner'
'cable'
'reserved'
'sans'
'presidential'
'russian'
'company'
'trump'
'monday'
'wednesday'
'political'
'2022'
'october'
'israeli'
'month'
'tuesday'
'attacks'
'democratic'
'donald'
'hezbollah'
'thursday'
'november'
'republican'
'iran'
'palestinians'
'media'
'july'
'facebook'
'putin'
'latest'
'2023'
'network'
'muslim'
'strike'
'party'
'chinese'
'politics'
'reuters'
'friday'
'sunday'
'hamas'
'invasion'
'fighting'
'2020'
'recent'
'government'
'rally'
'netanyahu'
'muslims'
'elections'
'islamophobia'
'ukrainian'
'june'
'moscow'
'killing'
'gaza'
'american'
'police'
'death'
'new'
'missiles'
'week'
'race'
'decades'
'jewish'
'conflict'
'rights'
'office'
'saturday'
'august'
'beijing'
'military'
'global'
'democrats'
'months'
'minister'
'bbc'
'fear'
'republicans'
'palestinian'
'fight'
'email'
'year'
'years'
'americans'
'twitter'
'islamic'
'2021'
'america'
'april'
'lebanon'
'instagram'
'september'
'injured'
'protests'
'february'
'old'
'violence'
'voters'
'weeks'
'terrorist'
'dead'
'2024'
'vote'

Streamlit App Dashboard



Dashboard Navigation

Select a section to view:

Dataset Overview

Preview Dataset

Rows per page

25

Sentiment Analysis Dashboard

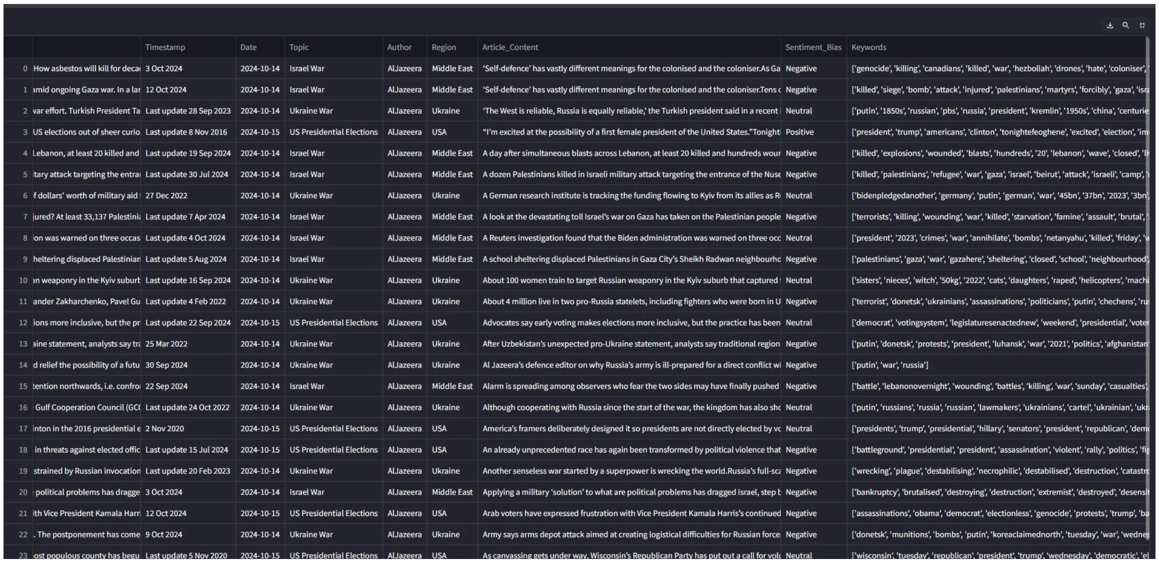
Using dataset: **Labeled_Dataset_with_Keywords.csv**

Overview Charts | Detailed Analysis | Word Cloud

Overview Charts

Overall Bias Distribution (Pie Chart)

Bias Distribution by Source



		Timestamp	Date	Topic	Author	Region	Article_Content	Sentiment_Bias	Keywords
0	How asbestos will kill for deca	3 Oct 2024	2024-10-14	Israel War	AlJazeera	Middle East	'Self-defence' has vastly different meanings for the colonised and the coloniser.As Ga	Negative	['genocide', 'killing', 'canadians', 'killed', 'war', 'hezbollah', 'drones', 'hate', 'coloniser',
1	amid ongoing Gaza war. In a lar	12 Oct 2024	2024-10-14	Israel War	AlJazeera	Middle East	'Self-defence' has vastly different meanings for the colonised and the coloniser.Tens c	Negative	['killed', 'siege', 'bomb', 'attack', 'injured', 'palestinians', 'martyrs', 'forcibly', 'gaza', 'isra
2	var effort. Turkish President Ta	Last update 28 Sep 2023	2024-10-14	Ukraine War	AlJazeera	Ukraine	'The West is reliable, Russia is equally reliable,' the Turkish president said in a recent	Neutral	['putin', '1850s', 'russian', 'pbs', 'russia', 'president', 'kremlin', '1950s', 'china', 'centurie
3	US elections out of sheer curio	Last update 8 Nov 2016	2024-10-15	US Presidential Elections	AlJazeera	USA	"I'm excited at the possibility of a first female president of the United States."Tonight	Positive	['president', 'trump', 'americans', 'clinton', 'tonightefeoghene', 'excited', 'election', 'im
4	Lebanon, at least 20 killed and	Last update 19 Sep 2024	2024-10-14	Israel War	AlJazeera	Middle East	A day after simultaneous blasts across Lebanon, at least 20 killed and hundreds wour	Negative	['killed', 'explosions', 'wounded', 'blasts', 'hundreds', '20', 'lebanon', 'wave', 'closed', 'li
5	itary attack targeting the entra	Last update 30 Jul 2024	2024-10-14	Israel War	AlJazeera	Middle East	A dozen Palestinians killed in Israeli military attack targeting the entrance of the Nuse	Negative	['killed', 'palestinians', 'refugee', 'war', 'gaza', 'israel', 'beirut', 'attack', 'israeli', 'camp', '
6	f dollars' worth of military aid	27 Dec 2022	2024-10-14	Ukraine War	AlJazeera	Ukraine	A German research institute is tracking the funding flowing to Kyiv from its allies as R	Neutral	['bidenpledgedanother', 'germany', 'putin', 'german', 'war', '45bn', '37bn', '2023', '3bn',
7	jured? At least 33,137 Palestini	Last update 7 Apr 2024	2024-10-14	Israel War	AlJazeera	Middle East	A look at the devastating toll Israel's war on Gaza has taken on the Palestinian people	Negative	['terrorists', 'killing', 'wounding', 'war', 'killed', 'starvation', 'famine', 'assault', 'brutal', '
8	ion was warned on three occas	Last update 4 Oct 2024	2024-10-14	Israel War	AlJazeera	Middle East	A Reuters investigation found that the Biden administration was warned on three occ	Neutral	['president', '2023', 'crimes', 'war', 'annihilate', 'bombs', 'netanyahu', 'killed', 'friday', 'w
9	heltering displaced Palestinian	Last update 5 Aug 2024	2024-10-14	Israel War	AlJazeera	Middle East	A school sheltering displaced Palestinians in Gaza City's Sheikh Radwan neighbourhc	Negative	['palestinians', 'gaza', 'war', 'gazahere', 'sheltering', 'closed', 'school', 'neighbourhood',
10	an weaponry in the Kyiv suburb	Last update 16 Sep 2024	2024-10-14	Ukraine War	AlJazeera	Ukraine	About 100 women train to target Russian weaponry in the Kyiv suburb that captured	Neutral	['sisters', 'nieces', 'witch', '50kg', '2022', 'cats', 'daughters', 'raped', 'helicopters', 'machi
11	ander Zakharchenko, Pavel Gu	Last update 4 Feb 2022	2024-10-14	Ukraine War	AlJazeera	Ukraine	About 4 million live in two pro-Russia statelets, including fighters who were born in U	Negative	['terrorist', 'donetsk', 'ukrainians', 'assassinations', 'politicians', 'putin', 'chechens', 'ru
12	ions more inclusive, but the pr	Last update 22 Sep 2024	2024-10-15	US Presidential Elections	AlJazeera	USA	Advocates say early voting makes elections more inclusive, but the practice has been	Neutral	['democrat', 'votingsystem', 'legislaturesenactednew', 'weekend', 'presidential', 'voter
13	aine statement, analysts say tr	25 Mar 2022	2024-10-14	Ukraine War	AlJazeera	Ukraine	After Uzbekistan's unexpected pro-Ukraine statement, analysts say traditional region	Negative	['putin', 'donetsk', 'protests', 'president', 'luhansk', 'war', '2021', 'politics', 'afghanistan'
14	d relief the possibility of a futu	30 Sep 2024	2024-10-14	Ukraine War	AlJazeera	Ukraine	Al Jazeera's defence editor on why Russia's army is ill-prepared for a direct conflict wi	Negative	['putin', 'war', 'russia']
15	tention northwards, i.e. confro	22 Sep 2024	2024-10-14	Israel War	AlJazeera	Middle East	Alarm is spreading among observers who fear the two sides may have finally pushed	Negative	['battle', 'lebanonovernight', 'wounding', 'battles', 'killing', 'war', 'sunday', 'casualties',
16	Gulf Cooperation Council (GC	Last update 24 Oct 2022	2024-10-14	Ukraine War	AlJazeera	Ukraine	Although cooperating with Russia since the start of the war, the kingdom has also sh	Neutral	['putin', 'russians', 'russia', 'russian', 'lawmakers', 'ukrainians', 'cartel', 'ukrainian', 'ukr
17	inton in the 2016 presidential e	2 Nov 2020	2024-10-15	US Presidential Elections	AlJazeera	USA	America's framers deliberately designed it so presidents are not directly elected by vc	Neutral	['presidents', 'trump', 'presidential', 'hillary', 'senators', 'president', 'republican', 'dem
18	in threats against elected offic	Last update 15 Jul 2024	2024-10-15	US Presidential Elections	AlJazeera	USA	An already unprecedented race has again been transformed by political violence that	Negative	['battleground', 'presidential', 'president', 'assassination', 'violent', 'rally', 'politics', 'fig
19	strained by Russian invocation	Last update 20 Feb 2023	2024-10-14	Ukraine War	AlJazeera	Ukraine	Another senseless war started by a superpower is wrecking the world.Russia's full-sc	Negative	['wrecking', 'plague', 'destabilising', 'necrophilic', 'destabilised', 'destruction', 'catastr
20	political problems has dragge	3 Oct 2024	2024-10-14	Israel War	AlJazeera	Middle East	Applying a military 'solution' to what are political problems has dragged Israel, step b	Negative	['bankruptcy', 'brutalised', 'destroying', 'destruction', 'extremist', 'destroyed', 'desensi
21	ith Vice President Kamala Harr	12 Oct 2024	2024-10-15	US Presidential Elections	AlJazeera	USA	Arab voters have expressed frustration with Vice President Kamala Harris's continued	Negative	['assassinations', 'obama', 'democrat', 'electionless', 'genocide', 'protests', 'trump', 'ba
22	. The postponement has come	9 Oct 2024	2024-10-14	Ukraine War	AlJazeera	Ukraine	Army says arms depot attack aimed at creating logistical difficulties for Russian force	Negative	['donetsk', 'munitions', 'bombs', 'putin', 'koreaclaimednorth', 'tuesday', 'war', 'wednes
23	ost populous county has begu	Last update 5 Nov 2020	2024-10-15	US Presidential Elections	AlJazeera	USA	As canvassing gets under way, Wisconsin's Republican Party has put out a call for volu	Neutral	['wisconsin', 'tuesday', 'republican', 'president', 'trump', 'wednesday', 'democratic', 'el