Russo-Ukraine War Analysis

Tommy Ngo, Cameron Zaidi, Revanth Rao
University of California, Davis
Dr. Peter Kramlinger

December 11, 2024

Abstract

The Russia-Ukraine War is a conflict of historic proportions, and has received significant political and media reactions. We use data from ACLED, SIPRI, and news articles from CNN and Fox News, with over 160,000 data points on conflict events and arms transfers, and thousands of media articles that were processed using API calls, dataset downloads, and web scraping. Key findings reveal Ukraine's reliance on U.S.-supplied anti-tank missiles and guided rockets, while Russia imported strategic missiles and aircraft from Iran and pre-invasion Ukraine. Conflict data highlights escalating battles and drone strikes starting in late 2023. Media analysis shows CNN's focus on advanced weaponry like drones and missiles, contrasting with Fox News's sporadic coverage. This study uses arms transfers, conflict data, and media framing to examine the conflict from the strategic, geopolitical, and narrative angles.

1 Introduction

Large-scale conflicts not only devastate the affected regions but also generate challenges that extend far beyond their borders, impacting global stability and economies. To address these crises effectively, it is crucial to analyze the types of military events, their geographical contexts, and how they are represented in media coverage. The media plays a critical role in shaping public opinion and sentiment, so it is important to evaluate the alignment of media narratives with on-the-ground realities.

The Russia-Ukraine conflict is the result of hundred of years of political tensions, but the crisis significantly escalated after Ukraine's independence in 1991, following the Soviet Union's dissolution. Relations between the two nations deteriorated in 2014, when Russia annexed Crimea after Ukraine shifted to a pro-European government. This caused violent conflicts in eastern Ukraine, with Russian-backed separatists fighting Ukrainian forces. The situation intensified further in February 2022, when Russia launched a full-scale invasion, triggering widespread devastation and a global reaction of economic sanctions against Russia and military aid to Ukraine.

This report examines the Russo-Ukraine War through a data-driven lens using armed conflict, weapons transfers, and media coverage data to analyze the war's events, political reactions, and news coverage over time. Through this analysis, we explore how the realities of the war evolved, and how the political and media responses differed as well. We hope this analysis can provide policymakers and leaders with the information needed to make better decisions on public policy, interventions, and peacekeeping.

2 Methods

2.1 Data Acquisition

This study draws on multiple sources, primarily the Armed Conflict Location & Event Data (ACLED) database, the Stockholm International Peace Research Institute (SIPRI), and articles from the Cable News Network (CNN) and Fox News. Each dataset provided unique insights into the Russia-Ukraine conflict, covering conflict events, weapons transfers, and media coverage. The following subsections outline the data collection and preprocessing procedures for each source.

2.1.1 ACLED Data: Conflict Events

The ACLED dataset was accessed through their Application Programming Interface (API), which allowed for efficient retrieval of conflict event data. We began by reviewing the API documentation to understand the available functionality, authentication requirements, and example queries. After requesting and obtaining an API key, we formulated queries based on parameters such as `country`, `region`, `event_date`, and `event_type`. These queries focused on the Eastern European region, specifically Russia and Ukraine, and covered the period from January 2022 to October 2024.

The API returned data in JavaScript Object Notation (JSON) format. To handle the large volume of data, we implemented a while loop to manage pagination, allowing us to iterate through multiple pages of results. Error-handling mechanisms were integrated to address potential issues, such as server downtime or invalid requests. Once retrieved, the JSON data was converted into a structured DataFrame using the `pandas` library. During this step, we performed initial cleaning, including handling missing values and standardizing column names for consistency. The resulting dataset, containing over 160,000 entries, was exported as a CSV file for further analysis.

2.1.2 SIPRI Data: Weapons Transfers

Unlike ACLED, the SIPRI dataset did not provide API access. Instead, we manually downloaded data from their website, applying filters for date range, region, and weapon supplier to focus on Russia and Ukraine. The data was provided in CSV format, which we imported into Python for preprocessing. This process involved cleaning missing values, normalizing column names, and standardizing formats to align with the ACLED dataset. The SIPRI data offered detailed insights into the diversity and volume of weapons transferred to both countries, serving as a critical input for our arms transfer analysis.

2.1.3 Media Data: CNN and Fox News Articles

In addition to the data from ACLED, we used the CNN and Fox News websites to web-scrape news articles about Ukraine. These articles were later used for keyword analysis to analyze the articles' sentiments. To get data from CNN, we used the search page on the website to find items containing the phrase "Ukraine". We then changed the filter to only include "Stories", which provided us only article results, we then used this link as our endpoint for the web scraping process in Python. Additionally, we explored the CNN search page using Developer Tools, which showed that the articles were grouped into pages of nine or ten articles each and that a "next" button was used to navigate to older articles. With this exploration complete, we began the web scraping by sending get requests to the CNN website, using Beautiful Soup to obtain the HTML and extract the link of each article. However, this process was unsuccessful as we were unable to extract the URLs in Python due to the containment of JavaScript in the HTML. To address this difficulty, we instead opted to use Selenium to scrape the article links. Selenium is a package created for web scraping and automating repetitive tasks in a browser. Using Selenium made our data collection significantly more difficult, as it required us to use a package not covered during discussion or on the homework assignments, and it increased the time required to execute the code. However, it did allow us to access the necessary articles by parsing the headline text element of the HTML. Then, we obtained the headline for each article using .text and the article links by obtaining the HTML for the "data-zjs-href" section. Afterward, we printed the page number of the CNN website that had just been scraped to verify that the process was successful, then paused the program from 1 to 3 seconds, randomly selected by Python before scraping the next page to avoid exceeding the rate limit.

With the collection of article URLs, we then created a function to extract the dates and article text for each link using Beautiful Soup. To understand the HTML structure of an article, we first examined one article using Developer Tools. Then, we began by sending a request to get the HTML for a given link, then extracted the date and cleaned it by removing the publishing time, leaving us with just the month, day, and year the article was published. Then, we used Beautiful Soup's find all function to extract all text in the paragraph element of the HTML, then cleaned the text to remove unnecessary elements like credit attribution or copyright symbols, leaving us with just the article text. Then, we added the dates and texts to lists, converted them to a Pandas data frame, and exported the 2,000 articles we extracted to a CSV file.

To gather Fox News articles related to the Ukraine-Russia conflict, we utilized the website's search function to filter relevant content. Upon initial inspection, the website displayed the first ten items in a structured list format, providing the title, date, a brief description, and a tag indicating whether the item was video or text-based; notably, most items were videos. At the bottom of the page, a "show more" button allowed us to load additional items.

We investigated whether the website preloads all articles by clicking the "show more" button to load additional content and noting a newly displayed article title. Using Beautiful Soup, we parsed the HTML of the original endpoint and searched for this title. Since it was absent from the HTML, we concluded that the website dynamically updates the page content upon clicking the "show more" button by requesting older articles from the server. Furthermore, when navigating back to the Ukraine-Russia page after clicking "show more," the articles reset to their initial state. This behavior suggests that the website does not use cookies or URL parameters to track the number of "show more" requests, complicating the scraping process.

To address this limitation, we employed Selenium to automate the loading of additional articles. To ensure proper loading, we created a loop to simulate clicking the "show more" button 3,000 times, with a two-second delay between clicks to allow for the website to fully load. We restricted the number of articles loaded due to limitations with both our browser and the Fox News server. We observed that attempting to load a significantly larger number of articles often caused the website to stall and revert to its original endpoint. This behavior could stem from several factors: the increased number of articles may strain both the user's system and the Fox News server, or Fox News may have implemented a cap on the number of articles a user can request. This process retrieved approximately 30,000 articles spanning 22 months. A get request was then sent via Python to capture the rendered HTML, which was then parsed using Beautiful Soup. Utilizing Developer Tools, we identified relevant HTML tags for extracting the title, date, and URL of each article. The data was then stored as nested dictionaries, a structured format which allows for efficient processing.

Upon inspection, we saw that video-based items contained the tag "video" in their URL, so we filtered out URLs that led to video content, thus retaining only text-based articles. For each valid URL, an HTTP GET request was issued to retrieve the article's HTML, and the text body was reconstructed by combining its paragraphs using methods that were similar to the CNN article processing. All extracted information was again organized into a Pandas data frame.

As part of preprocessing, we also standardized the publication dates, as articles published in less than 24 hours used a time-based format instead of a date. After doing this, the final dataset, containing 1,730 textual articles, was exported to a CSV file for subsequent analysis.

2.1.4 Data Integration and Preprocessing

To ensure consistency across datasets, we standardized date formats and column names during preprocessing, and resolved issues like duplicates and missing values. These steps facilitated smooth integration of the datasets, and allowed us to easily analyze the three topics at hand.

For topic analysis, we used the Latent Dirichlet Allocation (LDA) model, a probabilistic approach used to extracting themes from textual data. We preprocessed the data by removing non-alphabetical characters, converting words to lowercase, and tokenizing the text into individual words. We applied lemmatization to reduce words to their base forms (e.g., "running" became "run") and removed stop words that added no meaningful context, such as "said", "news", and "would." Afterwards, we vectorized the preprocessed data using Term Frequency-Inverse Document Frequency (TF-IDF) and fitted an LDA model. However, the results from the LDA model provided limited insights, so it was less effective for capturing themes within the articles.

Our keyword analysis provided more interesting insights. After following similar preprocessing steps, we identified articles containing these keywords: "drones", "transfers", "artillery", "missiles", and "strike".. We recorded their publishing dates, and plotted these keywords in a time series to examine trends in media coverage and analyze how news outlets shifted their coverage over the course of the conflict. This analysis was valuable in noting patterns in media narratives and how they meshed with real-world events during the war.

The data acquisition process gave us three comprehensive datasets: ACLED data containing over 160,000 data points of conflict events, SIPRI data with information on weapons transfers, suppliers, and weapon types, and CNN and Fox News data, which had a combined 3,730 articles detailing US media coverage of the war.

2.1.5 Challenges

Our project faced several significant challenges, highlighting the complexity of data acquisition and analysis in this study.

One major hurdle was accessing the SIPRI database. The absence of a documented API required us to manually download the data and perform extensive data cleaning and formatting to standardize the data for our analysis. This process significantly increased the time required for data acquisition.

The reliability of the ACLED dataset was a second challenge. ACLED compiles reports from various media sources and platforms like Telegram, so its datasets are could include instances of bias or incomplete reporting. This issue emphasizes the need for additional conflict datasets to ensure data accuracy.

Web scraping for CNN and Fox News articles also presented some technical difficulties. As shown in Figures 13 and 14, the websites for CNN and Fox News include a search bar followed by a list of articles, so using the Beautiful Soup package to scrape the website URLs was a natural first approach to collecting the articles. However, both websites were dynamically updated using JavaScript, so the links did not appear in the HTML obtained using Beautiful Soup. This issue necessitated that we use Selenium instead, and using Google Colab to run Selenium scripts required configuring a Chromium driver, which was both time-intensive and prone to various errors. Colab resource limitations added to the complications, as we had to handle interruptions and redo the web scraping a few times. Navigating JavaScript-heavy websites to extract article URLs while avoiding the rate limit added to these challenges, as this step added a significant amount of time to our data acquisition process.

Keyword analysis added limitations, as focusing on just terms like “drones”, “missiles”, and “artillery” may have excluded other potentially relevant phrases that discuss a different aspect of the war, such as “siege” or “counteroffensive”. Furthermore, the media can change the key terms they use in reporting, so simply focusing on a set group of words like “strike” may not capture the nuances of reporting. Approaches like using a larger keyword list, using natural language processing techniques to match synonyms, or periodically updating terms to reflect changes in news reporting can help with this issue.

Despite these obstacles, we were able to handle the challenges we faced in data acquisition to provide a thorough and insightful analysis of the Russo-Ukraine War.

2.2 Exploratory Data Analysis

2.2.1 Ukrainian Arms Transfers Analysis

Using data from the SIPRI dataset, we analyzed the types of weapons delivered to Ukraine and identified the key supplier nations, which are needed to examine political realities. Figure 1 shows the great volume of anti-tank missile and guided rocket transfers, with the United States as the major supplier. Figure 2 focuses on the financial aspect of these deliveries, presenting the top suppliers by Total Import Value (TIV). It again details the massive contributions of the United States, with Germany, Poland, and the United Kingdom also making significant contributions. Finally, Figure 3 explores the TIV of weapon types, revealing that while fewer tanks were delivered, they are more costly, which reflects their importance on the battlefield. From the figures, some of the most prominent weapons are various kinds of missiles such as guided rockets, surface-to-air missile (SAM), anti-tank missiles, beyond-visual-range air-to-air missile (BVRAAM), and anti-ship missiles (ASM). Military vehicle types such as Armored Personnel Carrier (APC) and Infantry Fighting Vehicles (IFV) were also common proceedings.

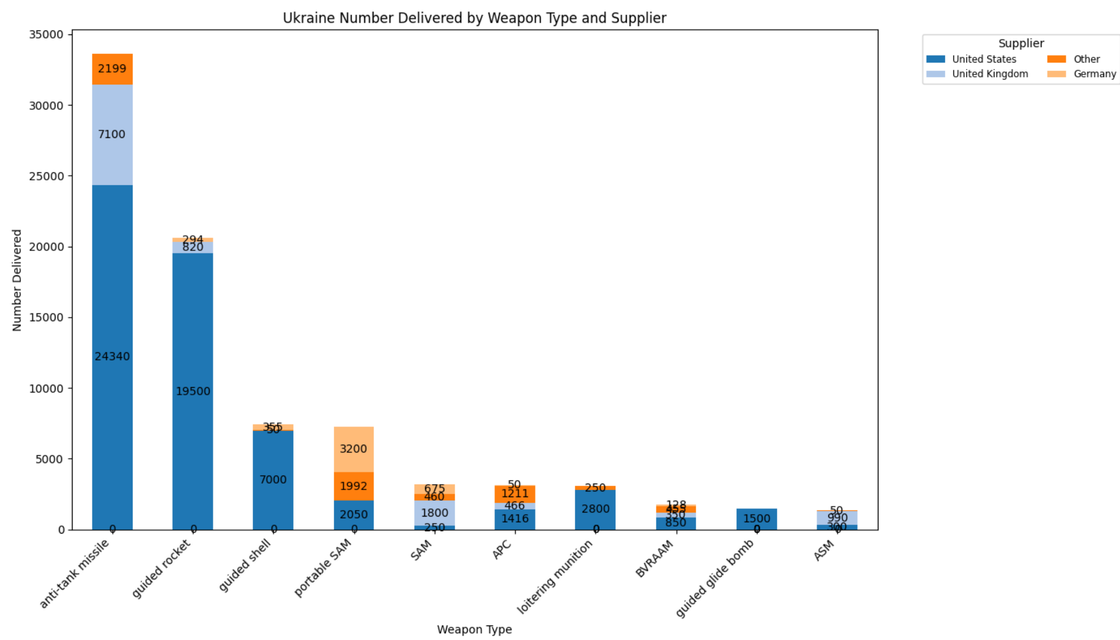

Figure 1: Number of weapons delivered to Ukraine by weapon type and supplier. Anti-tank missiles and guided rockets dominate the deliveries, with the United States as the primary supplier.

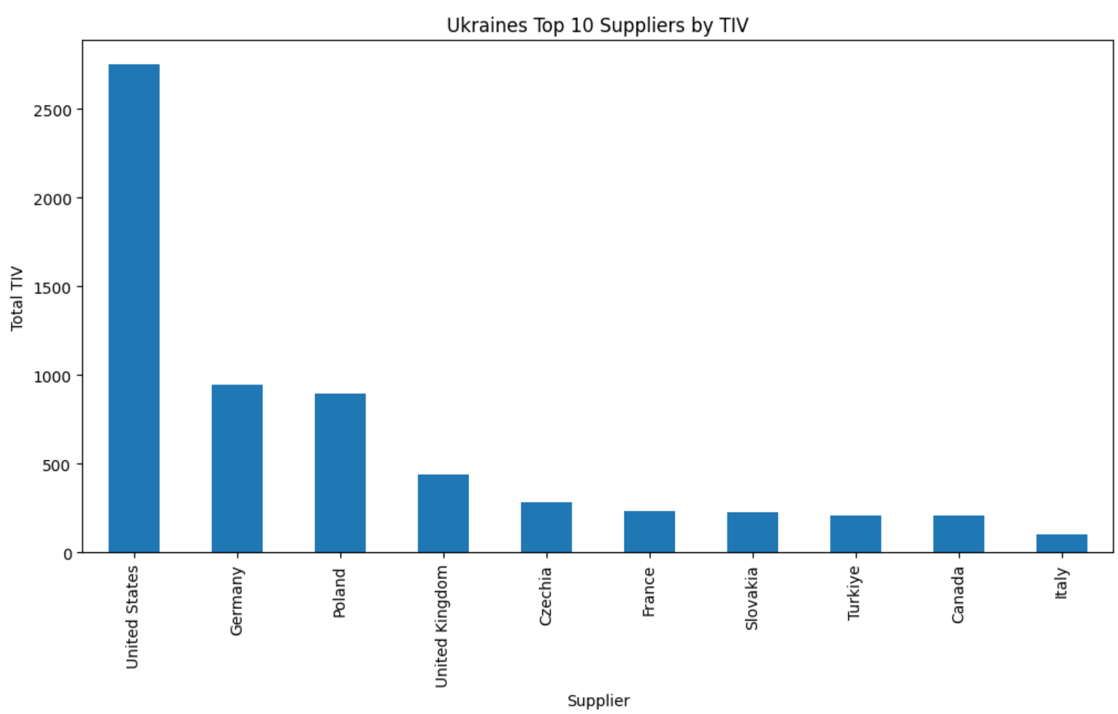

Figure 2: Top 10 suppliers to Ukraine by Total Import Value (TIV). It highlights the massive contributions of the United States, with Germany, Poland, and the United Kingdom also making significant contributions.

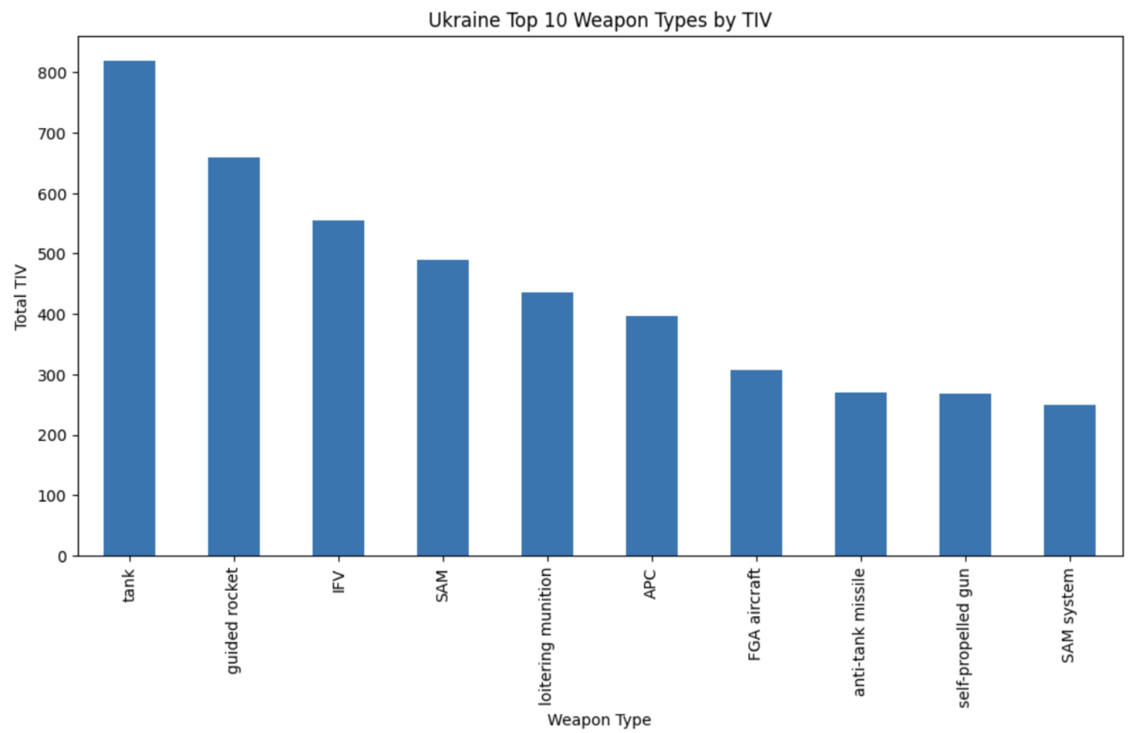

Figure 3: Top 10 weapon types delivered to Ukraine by Total Import Value (TIV). Tanks and guided rockets exhibit high costs, reinforcing their strategic importance despite lower transfer numbers.

2.2.2 Russian Arms Imports Analysis

Using data from the SIPRI dataset, we examined the weapon types imported by Russia and identified the top supplying nations. Figure 4 illustrates the number of weapons delivered to Russia, detailing a clear emphasis on Surface-to-Surface Missiles (SSMs) and Armored Personnel Vehicles (APVs) as the most imported weapons. Furthermore, the least common deliveries were Unmanned Aerial Vehicle (UAV) as well as Cargo Ships. Figure 5 analyzes the TIV of weapon types, detailing the prominence of transport aircraft and SSMs. Finally, Figure 6 emphasizes the top suppliers by TIV, showing Ukraine and Iran were significant suppliers to Russia. While it appears strange that Ukraine is the top supplier to Russia, this is due to a pre-invasion sale of costly AN-140 transport aircraft. Additionally, these figures show the diversity of weapons used by both countries and shows how political relationships influence which countries supply weapons to Ukraine and Russia.

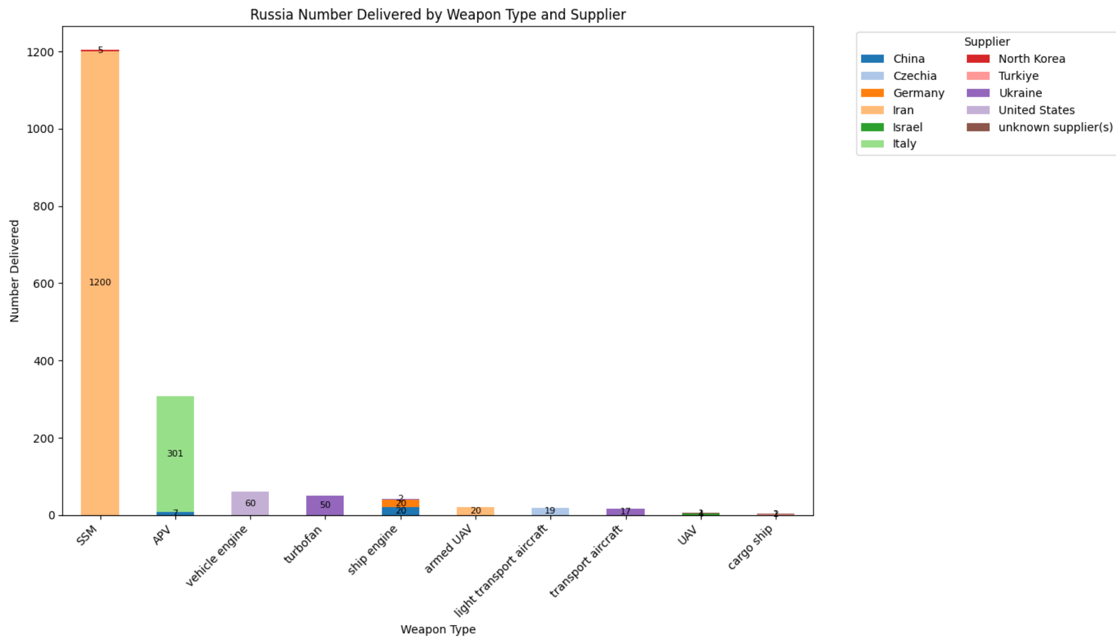

Figure 4: Number of weapons delivered to Russia by weapon type and supplier. SSMs dominate the imports, followed by APVs and various engines, highlighting a focus on strategic weaponry.

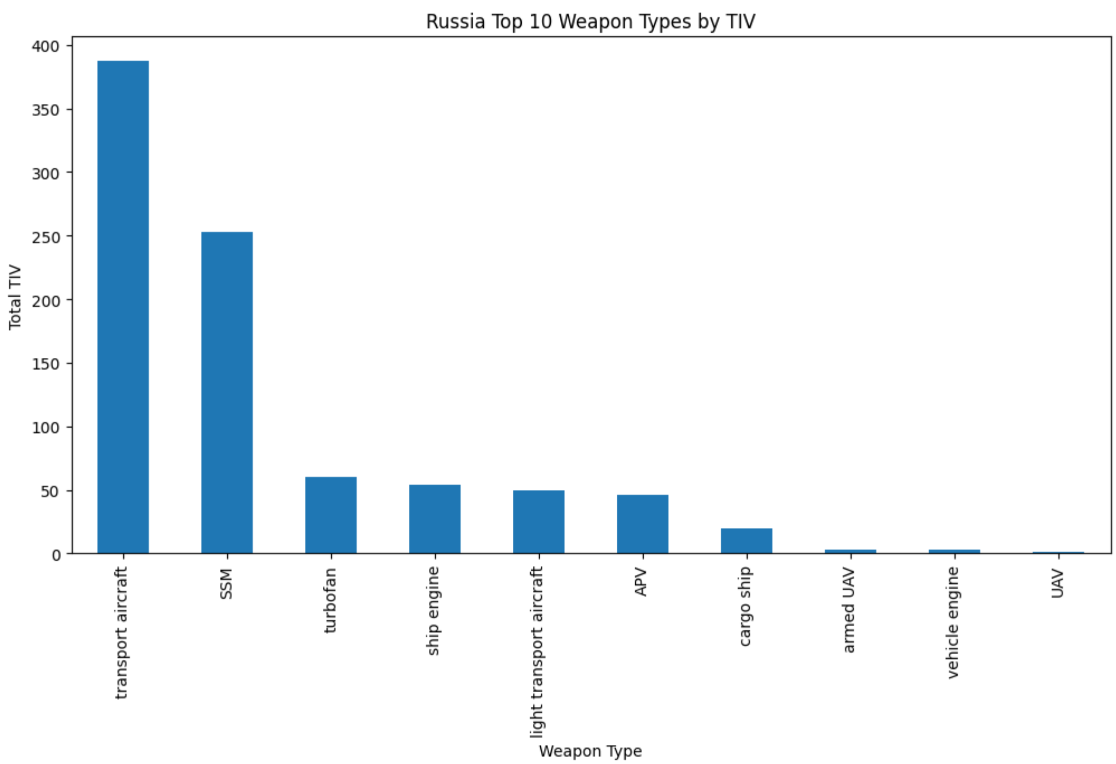

Figure 5: Top 10 weapon types delivered to Russia by Total Import Value (TIV). Transport aircraft and SSMs dominate in terms of financial importance, reflecting their strategic roles in the conflict.

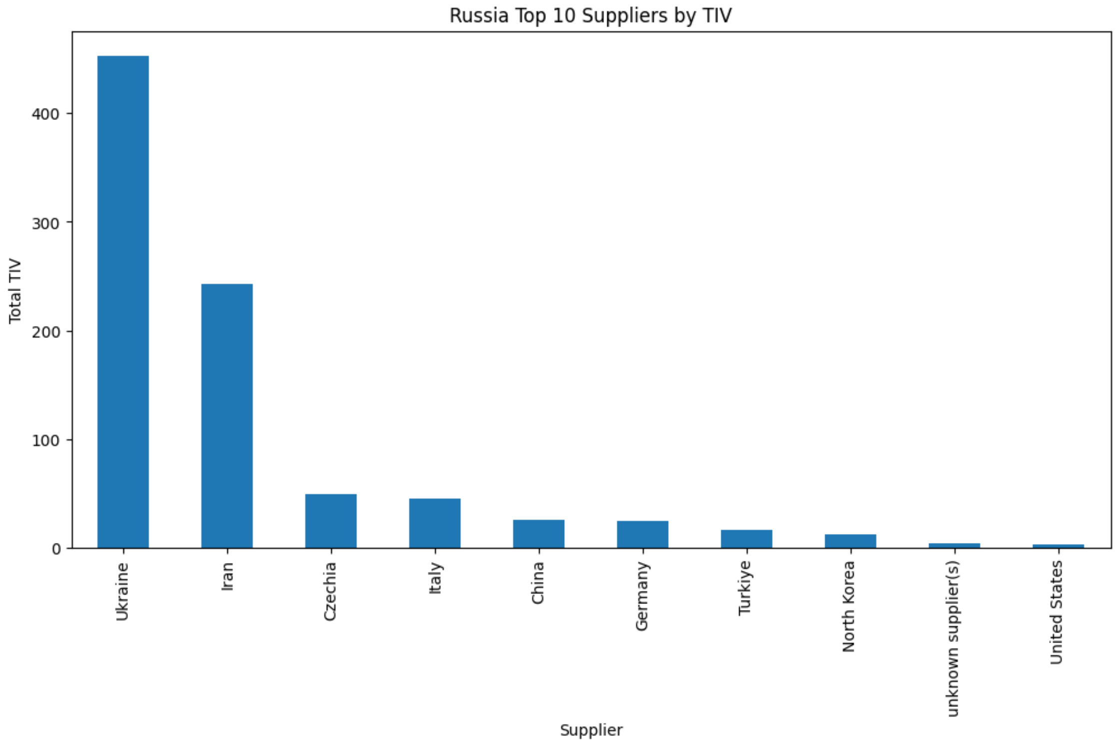

Figure 6: Top suppliers to Russia by Total Import Value (TIV). Ukraine and Iran lead as the primary suppliers, followed by Czechia, Italy, and China. While it appears strange that Ukraine is the top supplier to Russia, this is due to a pre-invasion sale of costly AN-140 transport aircraft[5].

2.2.3 ACLED Conflict Data Analysis

The ACLED dataset provides details the frequency and intensity of various events during the conflict in Ukraine. Figure 7 illustrates the frequency of different event types from 2022 to 2024. Explosions and remote violence are the most common event type, with battles becoming more frequent over time. Protests, riots, and strategic developments occur less frequently, but are still important in analyzing the conflict.

To provide a more specific view of certain event types, Figures 8, 9, and 10 analyze artillery strikes, armed clashes, and drone strikes individually using a time series and a 7-day moving average. It appears that shelling, artillery, and missile attacks increase initially, and remain relatively constant with peaks during offensives by Russia. Armed clashes appear to be steadily increasing, showing that the war has an greater emphasis on ground-level confrontations. The rise in air and drone strike events shows how critical these events are for remote violence and surveillance purposes.

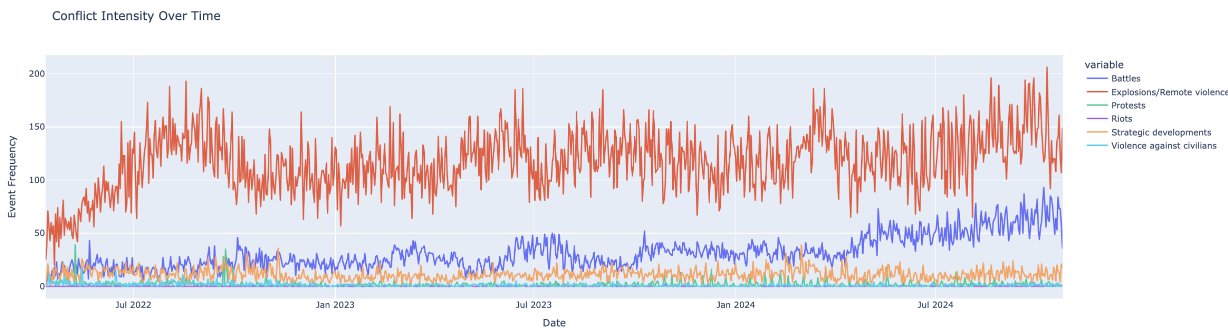

Figure 7: Conflict intensity over time, categorized by event type. Explosions and remote violence are the most frequent type, and battles are more common over time. Protests, riots, and strategic developments occur less frequently.

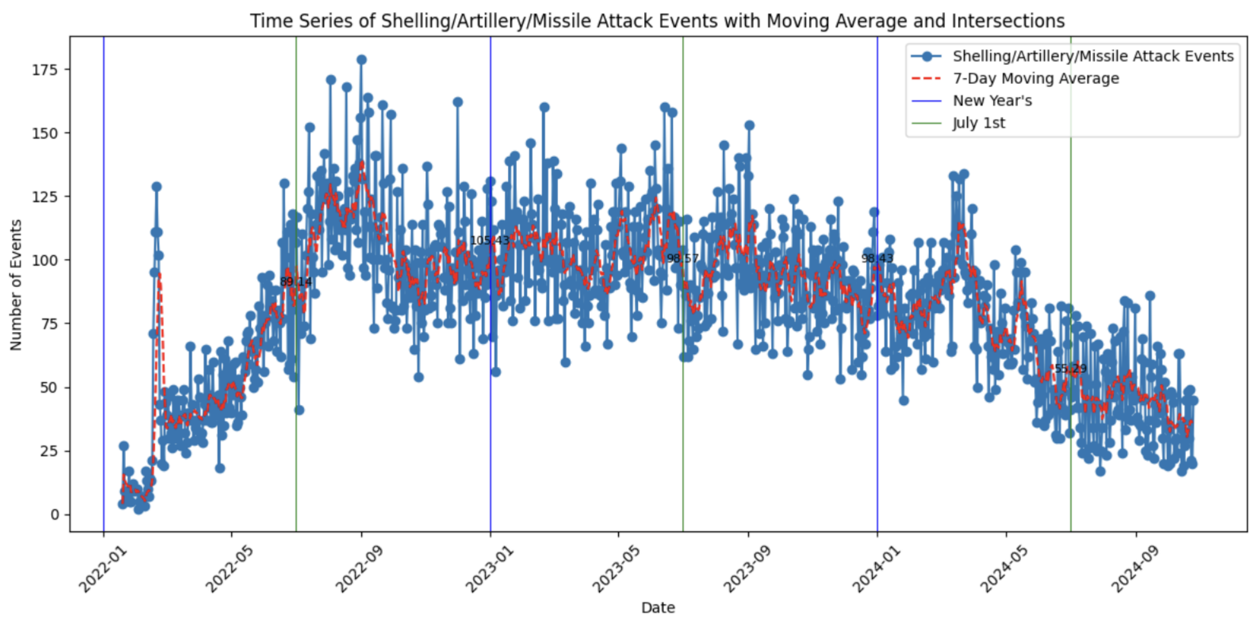

Figure 8: Time series of shelling, artillery, and missile attack events, with a 7-day moving average. Peak in the graph show military escalations and strategic offensives, and event frequency shows increases during mid-year.

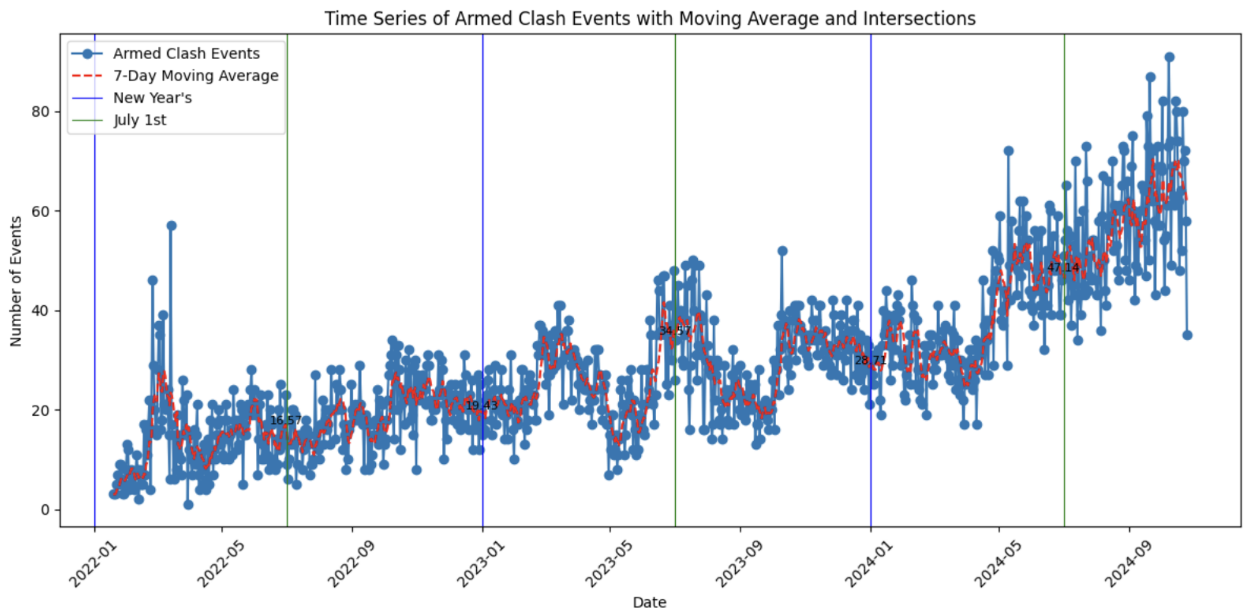

Figure 9: Time series of armed clash events, with a 7-day moving average. Armed clashes generally show a steady rise, especially in April of 2024, when President Putin announced a new offensive in Western Ukraine[6].

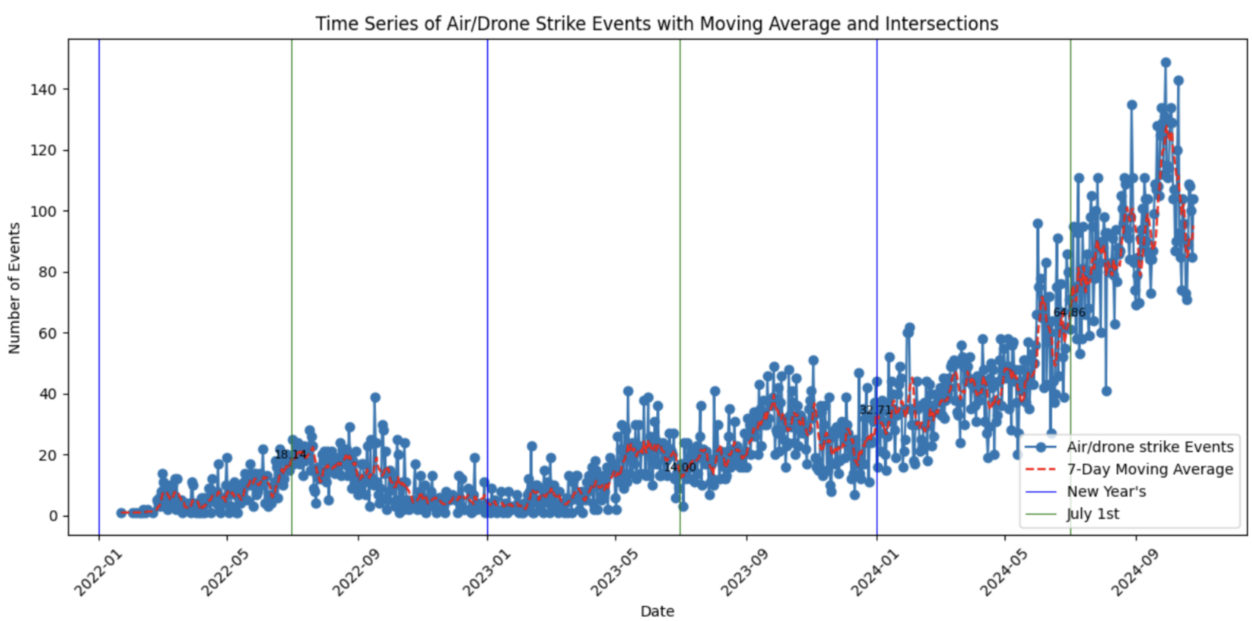

Figure 10: Time series of air and drone strike events, with a 7-day moving average. Drone strikes increase significantly over time, highlighting their importance in modern warfare and Russia's various offensives.

To add to this analysis, we developed an interactive map visualizing the geographic distribution of sub-event types across Ukraine. This map is available in the project repository for further exploration. The map enables users to filter by specific event types, such as missile attacks, air/drone strikes, protests, and territorial changes, offering a perspective on various conflict patterns. A screenshot of this interactive map is provided in Figure 11, highlighting the concentration of events in certain conflict zones and border regions, and highlights the important of studying the geographic spread and intensity of conflict events.

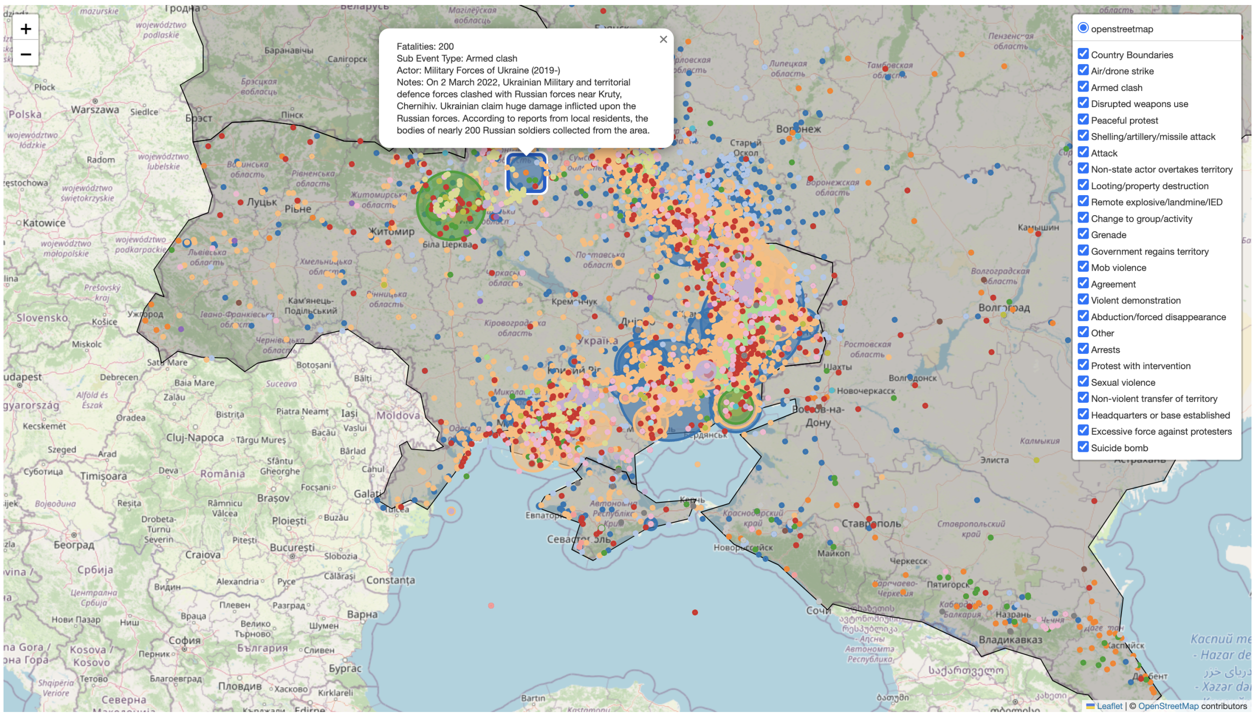

Figure 11: Screenshot of the interactive map created using ACLED data, visualizing the spatial distribution of sub-event types during the Ukraine conflict. The map includes filters for various event categories and demonstrates clustering in high-conflict areas.

2.2.4 Media Topic Analysis: CNN vs. FOX News Reporting

The analysis of media reporting provides valuable insights on how major news outlets discuss war events and political developments. Using CNN and FOX News articles, we tracked the frequency of key terms such as "drones," "artillery," "missiles," and "battles" over time. Figure 12 shows the frequency of six key terms over the course of the war. Based on the graphs and our previous analysis, it seems that the media accurately mentions drones and missiles more often during times of increased drone strikes and missile attacks, although CNN appears to mention these topics more often than Fox News. The same appears true for artillery and battles, as CNN focuses on these two sub events more than Fox. On the other hand, both Fox and CNN mention transfers more often when major announcements and deliveries occurred, although CNN appeared to have more coverage of these events sooner than Fox News. Both outlets also display greater keyword mentions beginning in 2023, but Fox News seems to have peaks after CNN, suggesting that Fox appears to have a more reactionary approach to reporting. In general, it appears that CNN simply uses more of these keywords than Fox News, indicating that CNN likely publishes stories with more detail about these keywords, while Fox appears to use a more general reporting style and focuses on other keywords as well. This is expected, as CNN and Fox News have differing ideologies and publish articles with disparate opinions on major events and political developments[7].

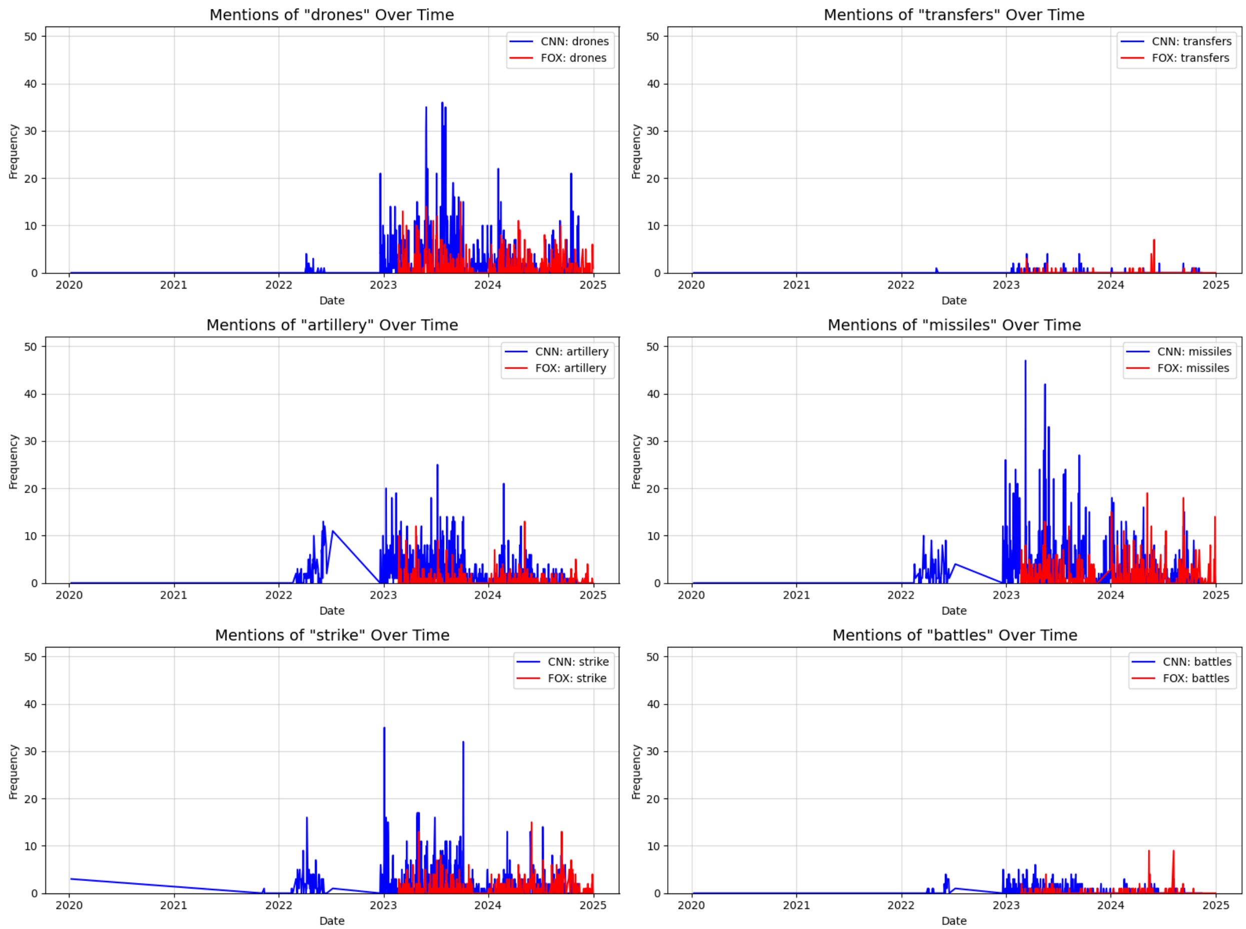

Figure 12: Mentions of key terms over time in CNN and FOX News. The frequency of mentions for terms such as "drones," "artillery," "missiles," and "battles" varies significantly between the two outlets, as the two outlets appear to have different priorities in reporting.

3 Conclusion

This project provides a clear analysis of the Russo-Ukraine War by integrating data from the ACLED database, SIPRI arms transfers, and CNN and Fox News articles. The time-intensive data acquisition process, consisting of API integration, manual downloads, and web scraping was crucial to the success of the project, and our data processing ensured the reliability and accessibility of the datasets. Using these datasets, we uncovered key insights on the geographic patterns of conflict events, the geopolitical nature of arms transfers, and variance in media coverage between major outlets. This approach highlighted the how military strategies, international support, and public narratives all mesh together and demonstrates how data-driven analysis can aid in understanding complex global issues.

Contributions

The contributions for this project are as follows:

- **Tommy Ngo and Revanth Rao:** Designed and implemented the web scrapers for CNN and Fox News using Selenium and BeautifulSoup. They played a critical role in ensuring the successful extraction of article data and contributed to writing, editing, and formatting the final report.
- **Cameron Zaidi:** Led the project and handled all remaining aspects, including data acquisition from ACLED and SIPRI, exploratory data analysis, creating visualizations, integrating media and conflict data analyses, and writing the report. Additionally, Cameron was responsible for organizing the code repository and creating the interactive ACLED sub-event map.

References

1. Armed Conflict Location & Event Data (ACLED). `https://acleddata.com/`
2. Stockholm International Peace Research Institute (SIPRI). `https://sipri.org/`
3. CNN. `https://edition.cnn.com/`

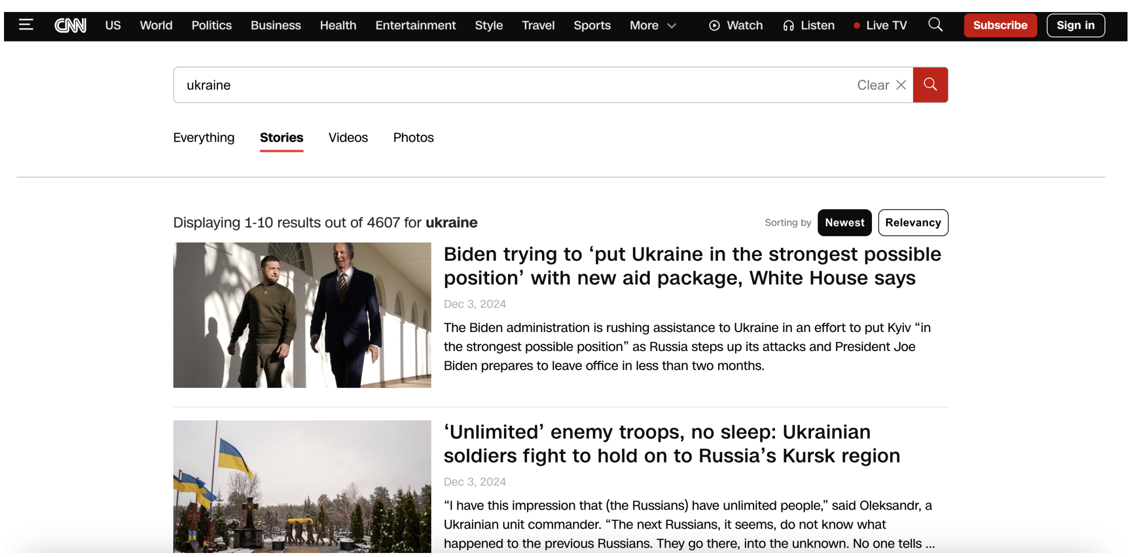

Figure 13: Search results on CNN's website for the term "Ukraine".

4. Fox News. `https://www.foxnews.com/`
5. Airforce Technology. `https://www.airforce-technology.com/news/news118549-html/`
6. Ukraine Conflict Tracker. `https://www.cfr.org/global-conflict-tracker/conflict/conflict-ukraine`

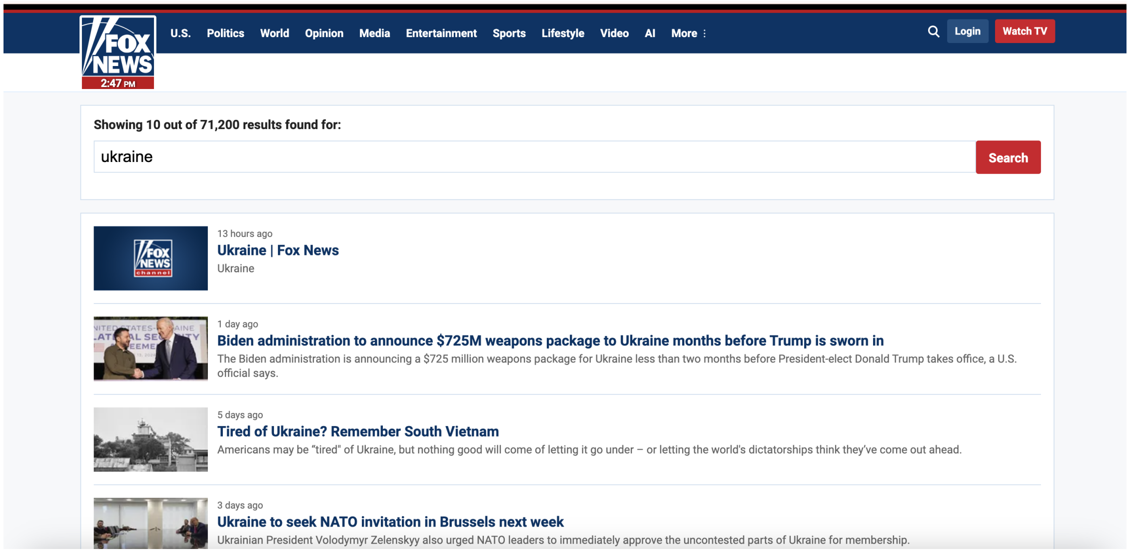

Figure 14: Search results on Fox News' website for "Ukraine".

7. Yang, J. (2023). How mass media influences U.S. political polarization— a comparison study of CNN and Fox News. *SHS Web of Conferences*, 178, 02005.

Code Repository

The complete code, data, and documentation for this project are available at the following GitHub repository: `https://github.com/camerontzaidi/Ukriane-Analysis/tree/main`

The repository includes the following components:

- `141B_project.ipynb`: A Jupyter Notebook containing all the code used for data collection, processing, analysis, and visualization. This notebook provides code used for data acquisition from ACLED API and the SIPRI datasets, webscraping from CNN and Fox News, data preprocessing, cleaning, and merging, visualizations for the datasets, and keyword-based sentiment and frequency analysis for news articles.
- `data/`: A folder containing some of the datasets used in the project.
- `README.md`: A detailed file documenting the project's objectives, methodology, and instructions for running the Jupyter Notebook. It includes setup steps, library dependencies, and guidance on reproducing the analysis.

The `141B_project.ipynb` notebook is self-contained, with markdown cells explaining each step of the process, making it easy to follow the workflow and replicate our results. The repository is designed to ensure transparency and reproducibility of the entire project.