Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 24, 2025
A Regional Conference of the Society for Industrial and Systems Engineering

Gaining the Edge: Visualizing Information Advantage through Machine Learning-Driven Dashboards

Ameir El Ouadi, William Knowlton, Adrian Pimentel, and David Beskow

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding Author's Email: knowltonw@gmail.com

Author Note: The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: The scale of online news data reveals opportunities and challenges for analysis and workflows across academic, commercial, and government sectors. This paper introduces an intelligent data system that provides competitive information advantage by leveraging the open-source news datasets of Common Crawl News (CC-News) and the Global Database for Events, Language, and Tone (GDELT). Starting with a robust data ingestion pipeline of metadata extraction and machine learning enrichment through Named Entity Recognition and language-agnostic sentence embeddings, our high-performance interactive visualizations capture and visualize key insights from the data at scale. The system offers efficient search and analysis of large multilingual news datasets and text corpora that exceed 250 million articles, pinpointing narrative, geographic, and network trends through custom analysis. This research enhances real-world decision-making by synthesizing multiple machine learning models to create robust information advantage dashboards from open-source news data.

Keywords: Open Source Data, News Data, Named Entity Recognition (NER), Common Crawl, GDELT, Machine Learning

1. Introduction

With the exponential growth and variety of online news, news data becomes increasingly valuable in describing people, places, events, and various narratives. Analysis of this data facilitates informed decision-making across business, social, political, government, and national security sectors. While this data can be crawled independently or purchased from a data-as-a-service (DAAS) company, two straightforward ways to access news data are through the Common Crawl News (CC-News) feed or the Global Database of Events, Language, and Tone (GDELT).

Our work advances news analysis by leveraging open-source data to deliver near real-time, multidimensional insights (narrative, geospatial, and network) within a single platform for applications in situational awareness, event summarization, and topic modeling to gain information advantage. We present an intelligent data system that operationalizes Common Crawl News and GDELT data for diverse academic, commercial, and government use cases. This paper introduces both Common Crawl News and GDELT data, reviews related research, and outlines a method for parsing, filtering, enriching with machine learning, modeling, and ultimately operationalizing this rich data stream for enhanced decision-making through dashboards.

2. Background and Previous Work

News narratives, or accounts of current and connected events, substantially motivate a society's cultural, political, and economic movements. In the context of information advantage, the ability to track narrative development, identify sentiment, relate key entities, and make predictions more comprehensively and rapidly than adversaries is a valuable asset. Although preliminary topic modeling finds roots in information retrieval (Blei, Ng, & Jordan, 2003), approaches evolved from Bayesian probabilistic models and Latent Dirichlet Allocation (LDA) (Zhao et al., 2021) towards artificial intelligence and large language models to detect and present narratives from text at scale (Pastorino, Sivakumar, & Moosavi, 2024).

Narrative visualization is seen in everyday life, from advertising and business intelligence reports to public service announcements. In news analysis, narrative visualization lets readers quickly identify key entities and events within a text corpus. Liu et al. (2012) broke ground through TIARA, an interactive, visual text summarization and analysis platform based on LDAs for the development of topic modeling (Liu et al., 2012). Over time, narrative visualization evolved with JavaScript D3 and other open-source tools, though it remained largely limited to highly structured data (Costa & Nunes, 2023). Raman et al.'s ViziTex (2021) supported high-performance visualizations of unstructured news data and text corpora at scale through hyper-

ISBN: 97819384969-9-8

graphs, radial layouts, topic graphs, and document hierarchy visualizations (Raman, Shah, Balch, & Veloso, 2021). Although such visualizations effectively communicate a narrative, they often lack the geographic context of their data.

Once news events are related to location, researchers can produce geographic relevance or patterns to their audiences. However, the challenges lie in enriching and extracting geospatial data from a text corpus. Successful methods include Entity Resolution (ER) to identify real-world geographic entities from their text mentions and natural language processing. However, disambiguation is still a chief concern (Balsebre, Yao, Cong, & Hai, 2022). Fortunately, the spread of online news and computerized mapping capabilities has rendered a robust collection of geographic news data. Geospatial information extraction and visualization methods have moved towards thematic maps, which portray a specific narrative theme with variation over the physical event space (Lawhead, 2013). Looking forward, innovations in machine learning will pave the way for emerging methods such as topical mapping, sentiment analysis, and networks of narrative and geospatial data.

Network analysis studies the interaction between nodes. As applied to news networks, nodes typically represent entities such as news outlets, articles, key actors, or events, while their edges capture citations, shared topics, co-mentions, or hyperlinks. In practice, semantic networking improved understanding of news framing (Jiang, Barnett, & Taylor, 2016) (Baden, 2018), conveyed sentiment (Danowski, Yan, & Riopelle, 2021), and topic evolution (Huang et al., 2022). Entity networks allow for topic modeling (Spitz, Almasian, & Gertz, 2019), news recommendation (Zhang, Yang, & Xu, 2021), and mapping global connections through knowledge graphs and named entity recognition (Bellamy et al., 2024). Lastly, link networks have modeled international news stream flow (Himelboim, 2010), news credibility (Borah, 2014), and online footprint of organizations (Fu & Shumate, 2017) through hyperlink connections that form the backbone of web navigation (Bizer, Heath, & Berners-Lee, 2023). Open source tools like Gephi, ORA, `networkx`, and `igraph` enable researchers to explore large-scale networks, uncover clusters of frequently discussed topics, and track the spread of narratives (Bastian, Heymann, & Jacomy, 2009). These tools and approaches provide rich insights into online news's interconnected and dynamic nature.

Based on our research, these previous endeavors highlight the need for a single platform capable of managing vast amounts of accessible, open-source, and unstructured data, incorporating narrative, geospatial, and network dimensions.

3. Common Crawl and GDELT Data

Understanding and analyzing news data at scale requires access to datasets that capture the underlying breadth of diverse global news trends, societal behaviors, and media narratives embedded in the raw text of news articles. CC-News and GDELT are widely used sources for such analysis. The Common Crawl corpus gathers, stores, and maintains an open Web ARCHive (WARC) file repository containing raw web page data, metadata extracts, and text extracts. With 3-5 billion pages of accessible internet crawled monthly through the Apache Nutch CCBot web crawler, Common Crawl offers a competitive open-source dataset for researchers, entrepreneurs, and developers. In 2016, the Common Crawl Foundation made a subset news stream available, CC-News. The GDELT project is an open repository of news data, self-described as "A Global Database of Society." GDELT collects broadcasts, print, and web news from across the globe, spanning over 100 languages and dating back to January 1979. Among other measures, GDELT specifically annotates the actors involved, their interactions, sentiments, event topics, as well as relevant geospatial data, which CC-News lacks.

Despite crawling a large portion of the internet, CCBot cannot access websites that block it in their `robots.txt` file. Sites such as Reuters and The Wall Street Journal, as well as certain nations, block the crawling of their news, creating gaps in the Common Crawl News data. However, it still makes a large portion of the news landscape easily accessible for analysis. In contrast, GDELT sources articles from a custom crawler, introducing a distinct set of news sources. A prominent limitation of the GDELT dataset is the absence of full-text articles, which necessitates re-hydrating the URLs to conduct critical analyses such as NER and sentiment analysis. Given the minimal content overlap between GDELT and Common Crawl, leveraging both databases is beneficial for capturing a broader and more diverse representation of the online news ecosystem and addressing potential gaps in coverage (El Ouadi & Beskow, 2024).

4. System Overview

The news data pipeline begins with parsing articles to extract key information, including the title, main text, authors, publication date, description, URL, and source domain, as represented in Figure 1. A language identification module is then applied to determine the primary language of each article. To ensure content appropriateness, the data undergoes filtering to remove articles containing Not Safe For Work (NSFW) material. Once preprocessed, the dataset is prepared for machine learning enrichment, a process we run on a server with 500 GB RAM, 24 cores, and an NVIDIA A4500 20 GB VRAM graphics card. Examples of this enrichment include Language-agnostic Bidirectional Encoder Representations from Transformers (BERT) Sentence Embeddings (LaBSE), which transform the article's text into numeric vectors in high dimesional Euclidean space,

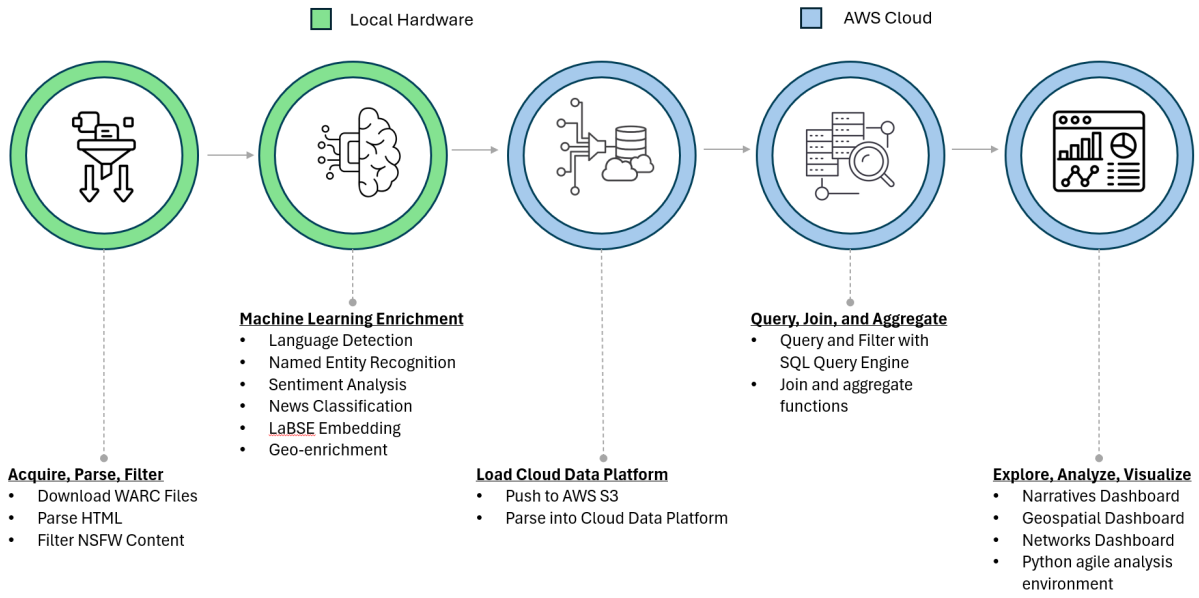


Figure 1: Information Advantage Dashboard Data Pipeline

enabling search and classification of articles. Sentiment analysis is also completed utilizing a pretrained model trained on robustly optimized BERT approach (RoBERTa) focused on 8 languages (Arabic, English, French, German, Hindu, Italian, Spanish, and Portuguese). This was trained on close to 200 million tweets and allows for capturing more nuances in language (Barbieri, Anke, & Camacho-Collados, 2022). Language Named Entity Recognition (NER) is also performed to extract entities such as persons, organizations, and locations. The pipeline geocodes location-based entities using a dictionary-based approach with a curated dictionary of worldwide location data down to city-level granularity. Following these enrichment steps, the processed data is ready for integration into various analytical tools, including knowledge graphs, language-agnostic search capabilities, and geospatially referenced maps, which can be incorporated into interactive dashboards. This system is described in greater detail in previous work (El Ouadi, Knowlton, Pimentel, & Beskow, 2025).

5. Methodology

Our system leverages a high-performance cloud data storage and querying platform that hosts large-scale datasets, such as Common Crawl and GDELT, providing comprehensive SQL query capabilities optimized for handling complex data structures. Leveraging SQL-based data engineering techniques enables precise filtering and extraction of targeted data subsets according to user-specified criteria. The platform currently holds 235 Gigabytes of data and 264 million records, 90% of which were produced by the pipeline described in this paper. Additionally, the platform supports advanced operations, including the unnesting and flattening hierarchical data structures, facilitating seamless integration into analytical pipelines. The refined datasets are visualized using interactive information advantage dashboards built with Apache Superset and shown in Figure 2. Apache Superset is an open-source business intelligence application designed to explore, visualize, and analyze data with a SQL query engine and interactive dashboards. Apache Superset enables insightful data-driven analysis of the world's news.

The News Narratives Dashboard offers a high-level analytical overview of the CC-News dataset, structured into two distinct tabs: one representing macro narrative trends across the user's query while the other explores named entities (which require advanced data manipulation) (Figure 2). The first tab comprehensively examines the CC-News data over the past two months (approximately 21 million articles), highlighting key metrics such as the most active source domains, article volume, and predominant publishing languages. The dashboard will also leverage a separate custom transformer model to enable the exploration of various article topics (i.e. Politics, Sports, National Security, etc).

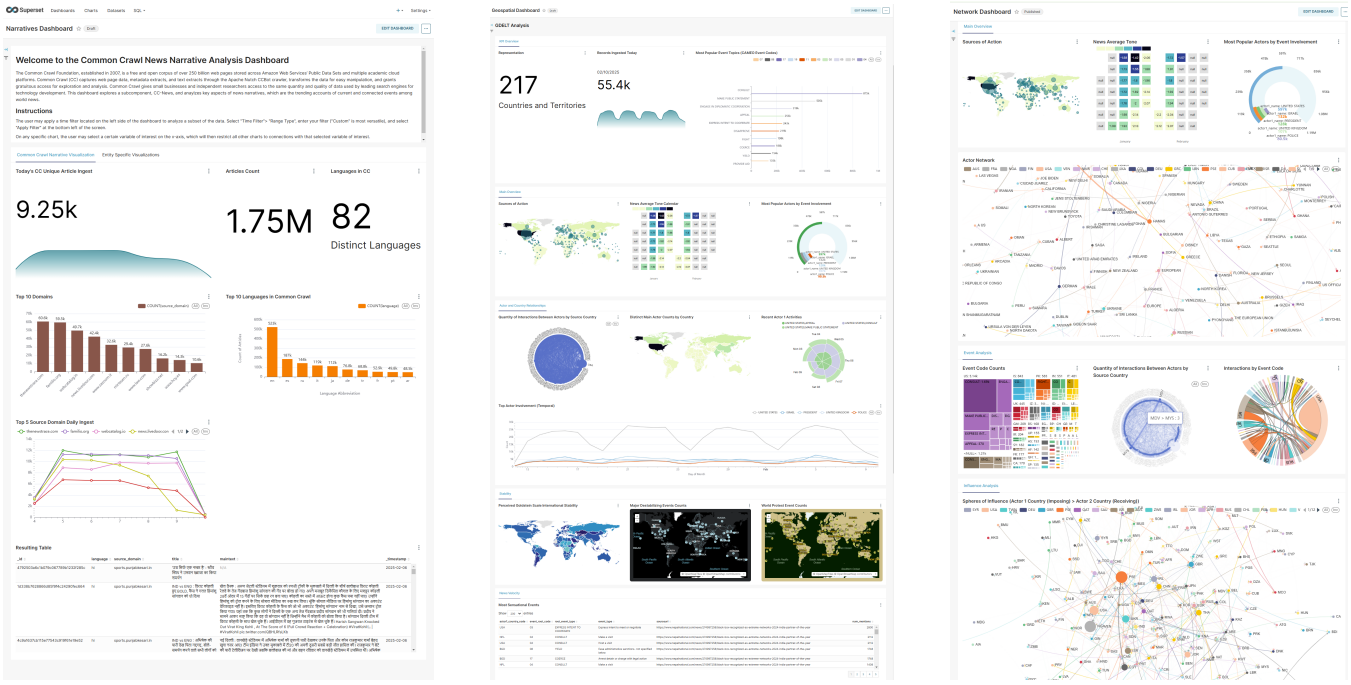
The second narrative tab explores named entities extracted from articles using pretrained named entity recognition (NER). The NER is conducted in seven language-specific models (English, Spanish, German, French, Russian, Portuguese, and Chinese) before running a language-agnostic model on all other languages. Any given news article contains many named entities. The underlying data pipeline will attach this list of named entities to the metadata associated with the article, creating a

nested data structure. A major part of our effort was creating the SQL data processes to "unnest" and display this named entities. This approach facilitates granular analysis and visualization of individual entities alongside their respective classifications (e.g., Person, Place, Organization). Visualizations within this tab include identifying the most frequently mentioned individuals, locations, geopolitical organizations, and products appearing in article texts. Both tabs incorporate filtering capabilities, allowing users to refine their analysis by language, time period, source domain, or entity type. Resulting tables of filtered results populate at the bottom of both tabs, allowing analysts to read individual articles.

Spatial context and relevance of current events are a critical base of knowledge for information advantage. The GDELT Geospatial Dashboard summarizes information by providing a main overview along with visualizations of Key Performance Indicators (KPIs) (Figure 2). It communicates the available data through article ingest timelines, the countries represented via event source mapping, news tone calendars, and the most popular event topics. The sub-groups focus on main actors and their activities over time, as well as event analysis through topic mapping, news velocity, world stability through Goldstein scale metrics (-10 to +10), and sentiment. The variables can be more effectively understood using the GDELT Event Database Data Format Codebook.

The News Networks Dashboard in Figure 2 complements the others by highlighting relational insights from the CC-News and GDELT datasets. It visualizes global interactions among actors, countries, and entities through dynamic, interactive network graphs. Users can explore intricate relationships using tools such as the actor network graph, which illustrates connections between individual actors; a chord diagram, displaying interactions categorized by event codes; visualizations depicting the quantity of interactions between actors by source country, and visualizations of author/source domain relationships, providing comprehensive insights into global connectivity and influence.

The primary limitations of the dashboards stems from the constraints of the Apache Superset software. Superset has a limited selection of visualizations, restricting the ability to incorporate more customized domain-specific graphics. This is particularly evident in the Networks Dashboard, where only a single network graph type can be displayed. Another drawback is the inability to conduct native keyword or substring searches in Apache Superset, requiring users to manually modify the upstream SQL query. The data platform does provide a JupyterHub environment that can conduct custom SQL queries and agile visualization, mitigating some of these limitations.



(a) Narrative Dashboard (b) Geospatial Dashboard (c) Networks Dashboard

Figure 2: Visualizing Common Crawl and GDELT News Data in Information Advantage Dashboards with Apache Superset

ISBN: 97819384969-9-8

6. Operationalizing Dashboards: A Congo Case Study

The mineral-rich regions of the Democratic Republic of Congo have been ravaged with conflict since the 1994 Rwandan genocide as armed assailants continue to compete for power (Zane, 2025). 26 January, 2025, marked the M23 rebel group's assault and capture of Goma, the North Kivu providence capital and vital logistical hubs (Zane, 2025). Despite being accused of providing direct support, Rwanda has denied any connection to the M23 rebel group (Zane, 2025).

Utilizing CC-News and GDELT, our dashboards provide analysis of the Congo events with depth, context, range of scope, and efficiency at scale. The CC-News narrative Dashboard rapidly orients the user to the quantitative summaries surrounding the events. The data was filtered from 5 January to 6 March 2025 to identify articles with titles containing "Congo". This resulted in 9,030 articles covering the event across 967 internet domains, peaking at 583 articles per day within the first 48 hours of the invasion. The bar charts show that events were reported in 17 distinct languages, most commonly English, French, and Spanish. Moving to the entity-specific analysis, the news covered 10.2k unique people, with Paul Kagame and Felix Tshisekedi, the respective presidents of Rwanda and the Congo, among the most common. 9.78k unique organizations and 929 unique geopolitical entities, mostly Congo, M23, Rwanda, Kinshasa, and Goma, represent the news coverage.

The GDLET Geospatial Dashboard provides context to physical event space of narratives. After filtering for the main actor country codes for Congo and Rwanda, our KPIs show that 19.3k articles covered the events, with the most popular topics being "consult," "make public statement," and "fight." Furthermore, the charts and maps show the most popular actors over time, with Rwanda and Congolese ranking highest, and Congo having 23 unique main actors. In terms of stability, our maps show that Congo had an average Goldstein stability score of -1.04 during this time period. Of all countries, Congo shows the worst average news sentiment in the world of -5.22. The News Average Tone Calendar nearly doubled in negative sentiment following the invasion on January 26th. Furthermore, we demonstrate news velocity through events similar to mentions, the most popular being Congo hosting a diplomatic visit on 31 January.

Lastly, the Networks Dashboard shows the interrelations between actors, countries, and other entities. With a network graph of source-to-target actors, the most significant edge is between Congolese and Rwanda. After filtering by event type "fight," we see a linkage of 37 events between these actors. Furthermore, in the source-to-target graph of the actor's home country, we can restrict either code to see which countries are influencing or being influenced through their quantity of interactions. The network dashboard also displays the most prolific authors and the respective source domains for which they publish. In our Congo example, Edith M. Lederer has written for over 16 domains regarding the invasion over the past 60 days.

The dashboards discussed above are highly interactive. Several global variables apply to specific datasets populating our charts. For example, users can filter the News Narrative dashboard with time ranges, main text sub-string searches, language, authors, domains, and titles. Time range, longitude and latitude, actor names, country codes, event topics and codes, sentiment, and similar mentions can filter the GDELT Geospatial Dashboard. Interactivity is further augmented with cross-filtering. Users can apply cross-filtering in our dashboards by selecting components in the Geospatial charts or entries in the Narratives table. This ensures that all other chart metrics update dynamically to display only the selected value's data. When generating dashboards specific to the Congo invasion, numerous insights can be gathered regarding the conflict due to this robust interactivity.

7. Conclusion

We present a robust, intelligent data system designed to operationalize Common Crawl and GDELT news data for various information advantage applications in academic, commercial, and government contexts. Our approach focuses on building an efficient data pipeline that parses and filters news articles, applies machine learning techniques like NER and language-agnostic sentence embeddings, and integrates the processed data into analytical tools for network analysis, geospatial analysis, and narrative visualization. Interactive dashboards enable users to gain actionable insights from large, multilingual datasets, empowering them to analyze trends, track key entities, and explore the global relationships within the news. Our system enhances the accessibility and usability of open-source news data, making it easier for users to derive insights from large and diverse news corpora. Implementing language-agnostic models and integrating geospatial data allows for more comprehensive analyses, regardless of linguistic or geographic boundaries.

Future work will focus on refining the system's ability to perform language-agnostic search, improving event/entity detection, and expanding the dashboard's visualization insights. We aim to increase the system's utility for decision-making in global contexts by further optimizing its scalability and enhancing its ability to handle various languages and cultures. This will provide significant information advantage to stakeholders across sectors and foster a deeper understanding of the narratives shaping our world.

ISBN: 97819384969-9-8

References

Baden, C. (2018). Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis. In *Doing news framing analysis ii* (pp. 3–26). Routledge.

Balsebre, P., Yao, D., Cong, G., & Hai, Z. (2022). Geospatial entity resolution. In *Proceedings of the acm web conference 2022* (pp. 3061–3070).

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2022). Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the thirteenth language resources and evaluation conference* (pp. 258–266).

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the international aaai conference on web and social media* (Vol. 3, pp. 361–362).

Bellamy, E., Farrell, K., Hopping, A., Pinter, J., Saju, M., & Beskow, D. (2024). Designing an intelligent system to map global connections. In *2024 ieee international systems conference (syscon)* (pp. 1–3).

Bizer, C., Heath, T., & Berners-Lee, T. (2023). Linked data-the story so far. In *Linking the world's information: Essays on tim berners-lee's invention of the world wide web* (pp. 115–143).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Borah, P. (2014). The hyperlinked world: A look at how the interactions of news frames and hyperlinks influence news credibility and willingness to seek information. *Journal of Computer-Mediated Communication*, *19*(3), 576–590.

Costa, M., & Nunes, S. (2023). Newslines: Narrative visualization of news stories. In *Text2story@ ecir* (pp. 37–46).

Danowski, J. A., Yan, B., & Riopelle, K. (2021). A semantic network approach to measuring sentiment. *Quality & quantity*, *55*, 221–255.

El Ouadi, A., & Beskow, D. (2024). Comparison of common crawl news & gdelt. In *2024 ieee international systems conference (syscon)* (pp. 1–3).

El Ouadi, A., Knowlton, W., Pimentel, A., & Beskow, D. (2025). Operationalizing common crawl news: Ai-enabled data pipeline for large-scale news analysis. In *2025 ieee international systems conference (syscon)* (pp. 1–3).

Fu, J. S., & Shumate, M. (2017). News media, social media, and hyperlink networks: An examination of integrated media effects. *The Information Society*, *33*(2), 53–63.

Himelboim, I. (2010). The international network structure of news media: An analysis of hyperlinks usage in news web sites. *Journal of broadcasting & electronic media*, *54*(3), 373–390.

Huang, L., Chen, X., Zhang, Y., Wang, C., Cao, X., & Liu, J. (2022). Identification of topic evolution: network analytics with piecewise linear representation and word embedding. *scientometrics*, *127*(9), 5353–5383.

Jiang, K., Barnett, G. A., & Taylor, L. D. (2016). Dynamics of culture frames in international news coverage: A semantic network analysis.

Lawhead, J. (2013). *Learning geospatial analysis with python*. Packt Publishing Ltd.

Liu, S., Zhou, M. X., Pan, S., Song, Y., Qian, W., Cai, W., & Lian, X. (2012). Tiara: Interactive, topic-based visual text summarization and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(2), 1–28.

Pastorino, V., Sivakumar, J. A., & Moosavi, N. S. (2024). Decoding news narratives: A critical analysis of large language models in framing bias detection. *arXiv preprint arXiv:2402.11621*.

Raman, N., Shah, S., Balch, T., & Veloso, M. (2021). Vizitex: Interactive visual sense-making of text corpora. In *Proceedings of the second workshop on data science with human in the loop: Language advances* (pp. 16–23).

Spitz, A., Almasian, S., & Gertz, M. (2019). Topexnet: entity-centric network topic exploration in news streams. In *Proceedings of the twelfth acm international conference on web search and data mining* (pp. 798–801).

Zane, D. (2025, February). *What's the fighting in dr congo all about?* Retrieved from `https://www.bbc.com/news/articles/cgly1yrd9j3o`

Zhang, X., Yang, Q., & Xu, D. (2021). Combining explicit entity graph with implicit text information for news recommendation. In *Companion proceedings of the web conference 2021* (pp. 412–416).

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.