A geolocated dataset of German news articles

Lukas Kriesch, Sebastian Losacker

Paper in Innovation Studies no. 2025/04

A geolocated dataset of German news articles

Lukas Kriesch and Sebastian Losacker

Lukas Kriesch (lukas.kriesch@geogr.uni-giessen.de, corresponding author)
Department of Geography, Justus Liebig University Giessen, Senckenbergstr. 1, 35390 Giessen, Germany

Sebastian Losacker (sebastian.losacker@geogr.uni-giessen.de)
Department of Geography, Justus Liebig University Giessen, Senckenbergstr. 1, 35390 Giessen, Germany; CIRCLE—Center for Innovation Research, Lund University, Lund, Sweden

Abstract

The emergence of large language models and the exponential growth of digitized text data have revolutionized research methodologies across a broad range of social sciences. News articles are an important source of digitized text data in this context. News data is crucial for the social sciences as it provides real-time insights into public discourse and societal trends, helping to understand various social phenomena and dynamics. However, most research involving news data is conducted at the national level, as geographically more granular news data is often unavailable. In this paper, we address this gap by providing insights into how news articles can be geolocated and how the texts can then be further analyzed. More specifically, we collect data from the CommonCrawl News dataset and clean the text data for further analysis. We then use a named-entity recognition model for geocoding, linking news articles to geographic locations. Finally, we transform the news articles into text embeddings using SBERT, enabling semantic searches within the news data corpus. In the paper, we apply this process to all German news articles and make the German location data, as well as the embeddings, available for download. As a result, we compile a dataset containing text embeddings for about 50 million German news articles, of which about 70% include geographic locations. The process can be replicated for news data from other countries, as we provide all code and workflows.

Keywords

News data, Natural Language Processing, Geography

JEL Codes

C55; C81; C45; R12; O33

Acknowledgements

We acknowledge financial support by the German Federal Ministry of Education and Research (BMBF 031B1281).

Background & Summary

The emergence of large language models (LLMs) and the exponential growth of digitized text data have revolutionized research methodologies across a broad range of social sciences.[1–3] The wealth of available digital sources has equipped researchers with the ability to conduct systematic, large-scale analyses through natural language processing (NLP) techniques. This shift has opened avenues for the development of innovative, text-based indicators that extend beyond traditional statistical metrics, providing timely, topic-specific, and geographically nuanced insights into social dynamics, public discourse, policy and economics.[4–6] The power of these indicators lies in their ability to reveal emerging trends, capture societal reactions to new developments, and highlight early signs of policy adoption or resistance, among many other fields of application. In geographical research, text-based analysis provides new insights into various phenomena at the local level, where established data sources often fall short. It offers an enhanced view of local, community, and regional dynamics, complementing traditional macro-level indicators by adding a more granular, context-specific layer of analysis. Recent studies have explored novel text-based data sources, including geolocated firm web pages, for regional analyses, but other types of geolocated text data have remained underutilised, leaving the rich potential of text data unexploited.[7–10]

While corporate and economic data provide structured insights, they often miss the broader cultural and social narratives that drive public opinion and collective behavior on the regional level. Researchers interested in studying social science phenomena from a geographical perspective therefore rely on additional data sources.

This is where news data becomes crucial: news articles embody the sentiments, perspectives, and legitimacy processes that underpin how events, policies, innovations, and societal changes are discussed, debated, and integrated into everyday life.[11–13] Such an “outside-in” approach offers researchers a unique window into public perception and the underlying forces shaping societal changes across various regions.

Applications of news data in the social sciences have contributed to fields such as economics [14–17] and political science.[18,19] However, much of this work has focused on macroeconomic or national-level trends. The potential to harness news data for more qualitative, regionally-focused insights remain underexplored.[20–23] News articles serve as a rich repository of local narratives, offering information on how regional identities, subnational dynamics, and cultural variations influence public acceptance and discourse around policies, innovations, and social movements, among other aspects.

In this paper, we present a comprehensive approach to analyzing large-scale news data by leveraging pre-trained transformer models, known for their exceptional semantic understanding. Focusing on the German subset of a global news corpus, we harness these capabilities to process and index extensive text data, enabling detailed insights into regional and thematic trends within German news coverage. Key to this methodology is the combination of data retrieval, advanced text embeddings, and semantic search, allowing for precise extraction and categorization of articles. Although the dataset described in this paper is limited to German news articles due to computational constraints and the feasibility of manual quality control, our approach can be adapted and replicated for processing news articles from other countries and in other languages. The paper's structure is designed to guide readers through the data processing workflow, from acquisition to practical applications, underscoring the importance of curated data handling and model adaptation for domain-specific research. The robustness of this strategy is verified through both qualitative analysis and quantitative metrics, emphasizing the model's adaptability and efficacy. The final dataset published with this article consists of an SQLite database and a Usearch vector database, which together provide comprehensive data storage and semantic search capabilities. The SQLite database contains structured information about news articles and their associated geographic locations, while the Usearch vector database enables efficient semantic search through vector representations of the articles.

Methods

Figure 1 illustrates the end-to-end processing pipeline for transforming raw web data into a structured dataset suitable for analysis. The process begins with downloading the dataset, followed by filtering out webpages with non-German country-specific top-level domains (TLD) to focus on relevant content. Afterwards, we extracted the primary text content from the raw HTML and retained only articles identified as German through language detection. Low-quality content was filtered out using established text quality heuristics.[24]
Next, we prepared training data for a custom named-entity recognition (NER) model to identify relevant entities. This model was used for comprehensive entity extraction and geocoding, linking entities to specific geographic locations. Finally, each news article text was transformed into a text embedding using SBERT, and these embeddings were stored in a vector database, allowing for efficient semantic search and downstream analysis. Detailed descriptions of each processing step are provided below.

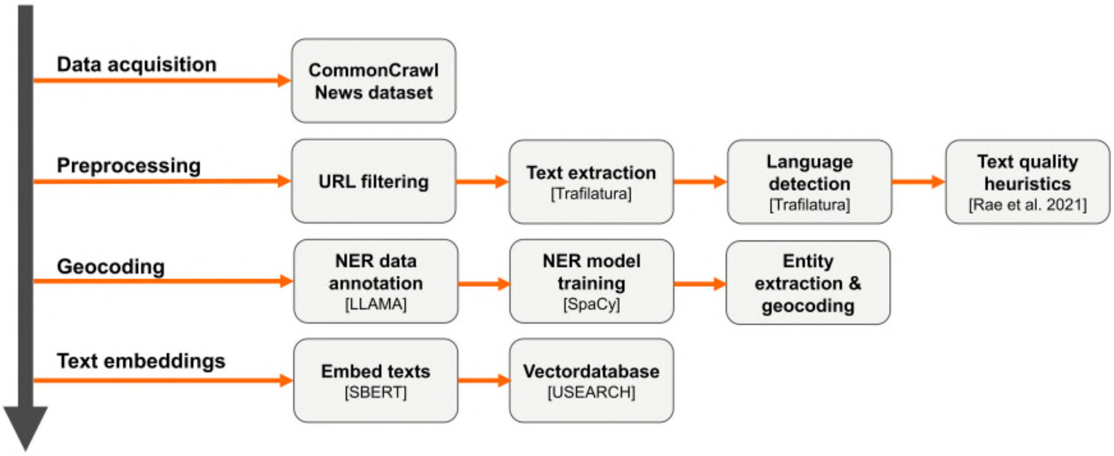

Figure 1: Data processing pipeline

Stage 1: Data acquisition

For data acquisition, we utilize the Common Crawl News dataset, a resource curated by the non-profit organization Common Crawl, which has been systematically crawling the web since 2007. This organization releases new collections of web content at intervals of 1 to 2 months. Since August 2016, Common Crawl has maintained a dedicated news dataset, using RSS/Atom feeds and news sitemaps to discover and aggregate links to articles across a wide spectrum of news platforms.

The dataset is provided in the WARC (Web ARChive) format, which includes both the complete HTML content of each webpage and metadata from the HTTP requests. This archival format enables researchers to access comprehensive records of web pages as they appeared at the time of crawling, which is essential for robust historical analysis and reproducibility in research.
Figure 2 depicts the size of each news crawl from August 2016 to December 2023 showing a consistent increase in the total compressed size of WARC files over this period. This growth trend highlights the substantial expansion of archived web data, with particularly rapid increases observed during and after 2020. The dataset used for this study encompasses over 35 TiB of uncompressed HTML text content, representing a vast source of raw news data.

Access to this data is facilitated through Amazon S3 buckets or by direct download from the Common Crawl servers, allowing flexibility for both cloud-based and local data processing.

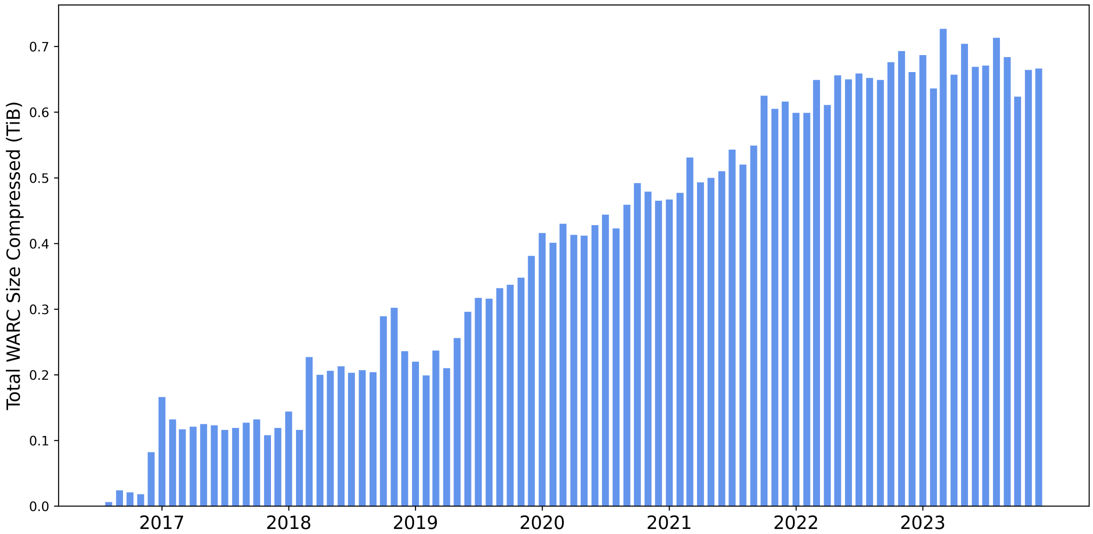

Figure 2: Size of news data collected by Common Crawl over time

Stage 2: Text extraction and pre-processing

The second stage of the pipeline involves filtering, extracting and preparing the raw HTML content from the Common Crawl News dataset for downstream analysis. Given the dataset's scale and heterogeneous sources, this stage is critical for transforming unstructured web data into clean, structured text suitable for text analysis and machine learning tasks.

In a first step, we removed entries with non-German country-specific TLD. This targeted filtering reduces the dataset's volume by eliminating sources that are unlikely to provide German-language content, thereby streamlining the subsequent language detection and quality filtering processes.

To ensure the relevance and quality of the extracted content, we eliminated extraneous elements such as navigation menus, headers, footers, boilerplate text, and advertisements. We employed the Trafilatura library for this purpose, which efficiently extracts not only the main body of the article but also associated metadata including titles, tags, categories, and excerpts when available.[25] Additionally, we utilized Trafilatura's built-in language detection algorithm to extract only German-language texts. Trafilatura has been validated as a fast and reliable tool for text extraction from web pages, offering significant improvements in text quality and accuracy.[26,27]

Web text data often contains substantial amounts of low-quality or poorly formatted content, which can introduce noise and diminish the accuracy of subsequent analyses. To mitigate this issue, we applied a rigorous filtering process based on established text quality heuristics.[24,28] We removed articles with five or fewer sentences to eliminate content that might lack depth and context, often appearing as stubs or summaries. Articles with more than 10% non-alphabetic words were also filtered out, as this could indicate a prevalence of numbers, symbols, or code snippets instead of narrative text. Additionally, we excluded articles averaging five or fewer words per line to avoid lists, tables, or poorly formatted content that does not resemble standard prose. We excluded articles containing JavaScript code. We retained articles with an average word length between 3 and 10 characters to ensure the language is typical, avoiding overly technical jargon or overly simplistic words. We removed duplicate articles by retaining only unique combinations of text and news provider. Lastly, we kept articles with a word count between 50 and 10,000 to exclude those too short to be informative and those excessively long, which might not represent genuine news content. This approach enabled us to systematically identify and exclude substandard data, ensuring that our corpus consisted of high-quality, reliable texts. After completing these preprocessing steps, the dataset comprises 49,374,999 German news articles.

Stage 3: Named entity recognition and geocoding

The geocoding of news articles is an important step for geographical analyses of news, and there are different ways in which a news article can relate to a specific geographical area or place. When analyzing news articles from a spatial perspective, it is essential to recognize that news is more than a mere collection of isolated events. News is created by individuals (I) to inform others (II) about events occurring in specific locations (III).[20] (I) The location of production represents where journalists and editors craft the content, potentially shaping how stories are framed and selected. (II) The consumption location reflects where the news is consumed, which can shape local opinions and decision-making. (III) Lastly, the event location indicates the geographic context of the incidents being reported, offering collective insights and perceptions of what is happening in a given area or place.

These spatial dimensions—production, event, and consumption—are critical for understanding how news circulates across regions and how it influences sentiments and perceptions in different areas. While detailed data on readership is often scarce, we focus in this paper on the relationship between locations and events, aiming to understand how regional narratives and discussions around specific topics develop and vary. To effectively analyze these spatial dimensions, particularly the event locations within news articles, it is essential to accurately identify and extract geographic entities. We employed a multi-step NLP pipeline to extract location entities from a large corpus of German news articles. Initially, we utilized Meta's LLAMA-3.1-8B-Instruct model for generating entity extraction responses. To process the text data, we sampled and deduplicated articles, yielding a set of 50,000 unique texts for analysis. Each text was inputted into the model along with a predefined system prompt designed to elicit structured JSON responses containing identified location entities. To ensure that the extracted entities met the expected types we validated them using Pydantic models.[29] Following validation, the entities were incorporated into spaCy's processing pipeline to train a custom Named Entity Recognition (NER) model specifically designed to identify city names within the news articles. The use of place names as point data enables precise assignment to territorial and statistical units. As a result, federal state and district names are excluded. Similarly, landmarks, street names, and square names are omitted, as their clear allocation cannot be guaranteed.

We chose spaCy for its fast CPU inference and lightweight model architecture, which is ideal for deploying efficient NER systems at scale. The performance metrics of this custom NER model are detailed in the technical validation section.

Using the LLM for annotation provided substantial benefits in scalability and consistency, as manually annotating such a large dataset would have been impractically time-consuming and resource-intensive. We considered the potential for biases in the LLM's annotations, which could reflect its training data. Notably, since the Common Crawl dataset is part of the LLM's training corpus, we infer that the model has robust knowledge of named entities in news articles. Manual inspections of the training data did not reveal any significant biases.

Despite these limitations, the annotations provided by the LLM served as high-quality training data for our custom NER model. This model achieved strong performance metrics and significantly reduced inference time, a crucial factor for handling large datasets. We utilized the model to extract location entities from the entire database of articles. To ensure consistency, we normalized the extracted location entities by converting them to lowercase, removing special characters, and eliminating any extraneous white spaces. Additionally, we excluded locations with fewer than 100 occurrences to reduce the likelihood of false classifications. Our analysis revealed that 36,305,239 articles, i.e., about 70 % of the whole news article corpus, contained at least one valid city name.

We used the normalized location names as input for the Nominatim geocoding service to obtain precise geographic coordinates.[30] These coordinates were then aligned with a NUTS-3 shapefile of Germany, allowing us to associate each location with its corresponding administrative region.

Stage 4: Embedding transformation

To facilitate semantic analysis of the geocoded data, we embedded the news articles using sentence transformers, converting the text data into numerical vectors. We encoded the articles with a "passage: " prompt, ensuring that the model treats each article as a discrete textual passage. This approach prepares the passages to be found effectively via semantic search, enhancing retrieval relevance and specificity for nuanced content queries. We employed the “deepset-mxbai-embed-de-large-v1” model to embed the texts.[31] This model facilitates Matryoshka Representation Learning and Vector Quantization, which effectively reduces memory consumption when analysing the data at scale. Additionally, we offer the embeddings in different vector dimensions and precisions, enabling semantic search at varying levels of precision and compatibility with diverse hardware requirements. We detail the usage of the vector database for article search and retrieval in the data records section.

Data description

This section provides an overview of the geolocated dataset. The complete published dataset includes all articles, including those without linked geographic information.

Article Description

Figure 3 illustrates the temporal distribution of news coverage per month, revealing a steady increase in the number of articles since the commencement of crawling in August 2016, with a peak reached in early 2020 during the COVID-19 outbreak in Germany. Since then, the number of articles per month has stabilized at around 500,000, with only minor seasonal fluctuations.

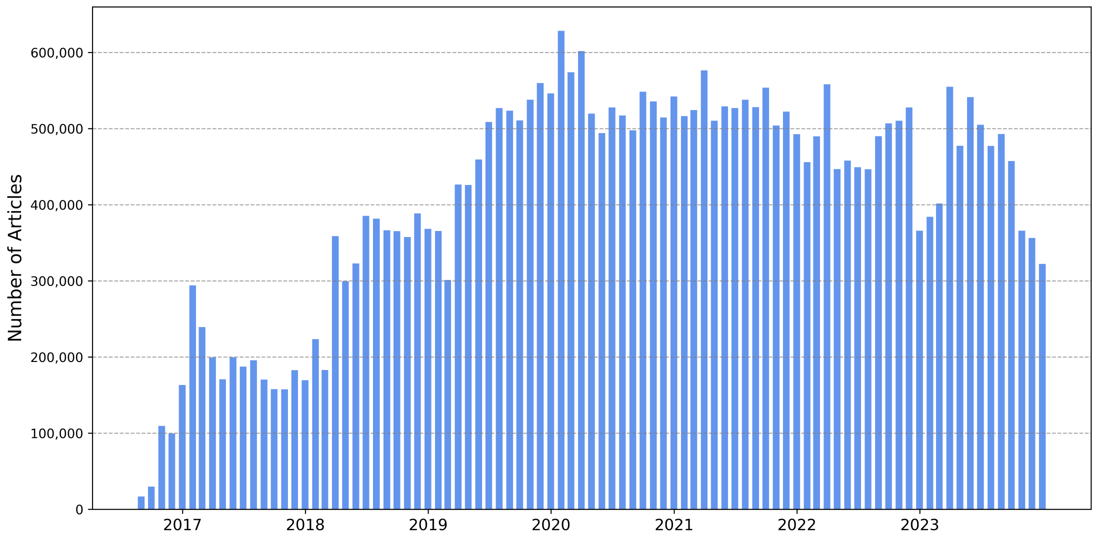

Figure 3: Number of German news articles per month

The database encompasses news items from 3,211 different domains. The Gini coefficient of 0.95 for the distribution of news articles across domains in Germany underscores a significant concentration of news media production, with a few large providers dominating the landscape. Figure 4 displays the lorenz curve of the distribution. Table 1 depicts the ten largest news sources with a strong concentration in finance-focused providers, led by *aktiencheck.de* and *finanznachrichten.de* with over 3.6 million. Major national outlets like *welt.de* and *presseportal.de* each contribute over a million articles, offering broad coverage across societal issues. Moreover, regional sources such as *augsburger-allgemeine.de* and *schwaebische.de* are also represented and capture local narratives that complement the national perspective.

News provider (hostname)	Number of articles
aktiencheck.de	3,659,083
presseportal.de	1,168,982
welt.de	1,167,094
finanznachrichten.de	869,548
stern.de	793,142
augsburger-allgemeine.de	719,559
schwaebische.de	598,506
rp-online.de	576,477
merkur.de	486,915
volksstimme.de	460,114

Table 1: Ten largest news providers

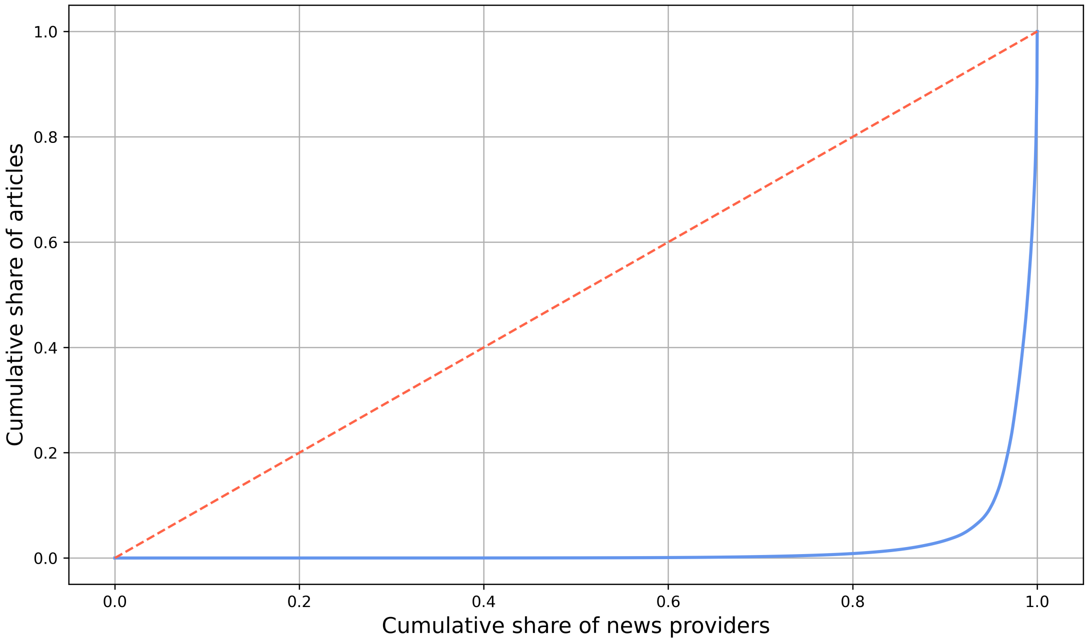

Figure 4: Lorenz curve of the distribution of German news articles across domains

This high level of skewness indicates that, although numerous media outlets exist, the majority of news content is generated by a small number of major players. This concentration results from the market dominance of these major outlets, which possess the resources and infrastructure necessary to produce and distribute news on a larger scale. However, the dataset also includes news from many smaller outlets, ensuring a diverse range of perspectives and regional coverage. This diversity is crucial for capturing the nuances and variations in news reporting across different regions.

The news articles analysed in this study are predominantly full text, averaging 410 words, as shown by the red dashed line in Figure 5. Relying on full text, rather than just headlines, is critical to grasping the depth and complexity of each article's narrative. While headlines are effective at capturing attention, they often prioritise engagement over nuance and can potentially skew the underlying message unintentionally. By analysing full texts, this dataset ensures a more robust representation of news narratives and sentiments, capturing subtle shifts in discourse that might otherwise be overlooked.

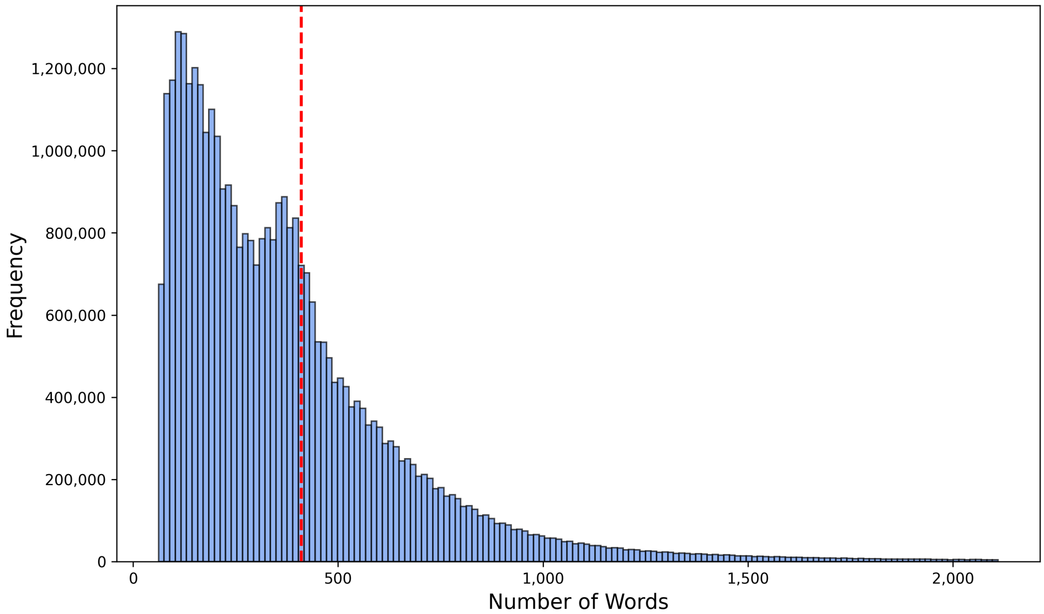

Figure 5: Word count per article (top and bottom 1% excluded)

In the dataset, 97.5% of the articles include an excerpt, providing readers with a brief summary to help them quickly grasp the main points. Nearly all articles (99.9%) come with a title, offering a clear initial indication of the article's subject matter and drawing readers in. The organization of articles is further enhanced by the use of tags, which are present in 64.88% of the articles, allowing for better categorization and easier navigation. Additionally, 32.13% of the articles are grouped into categories, further aiding in thematic analysis and organization.

Location analysis

We found 18,536 unique German locations in the dataset. The map in Figure 6 displays the number of news articles per location (log-scaled) using a hexbin density visualization. Each news article is assigned to a hexagon with an area of 44 km^2 based on the identified location(s). This type of visualization enables the assessment of count data on a map, illustrating the spatial distribution of news articles mentioning locations. It reveals spatial concentrations in the distribution of news mentions across Germany. Major cities such as Berlin, Hamburg, Munich, and Frankfurt dominate the news landscape, each being mentioned in a significant number of articles. One important observation from the map is that our dataset includes news articles associated with locations across the entire country, with only a few empty spots in sparsely populated areas of Germany. These empty hexagons predominantly cover forests, agricultural areas, or other unpopulated areas. When aggregating the location data to the NUTS3 level, the dataset includes news articles for all 400 NUTS3 regions. This comprehensive coverage enables regional analyses across the entire country, highlighting the dataset's value for geographical research. The map also clearly shows that the number of news articles per location correlates positively with population size, as evidenced by the high number of articles for major cities. This serves as a quality indicator for validating the accuracy of our geocoded data, a point that is discussed in greater detail in the Technical Validation section.

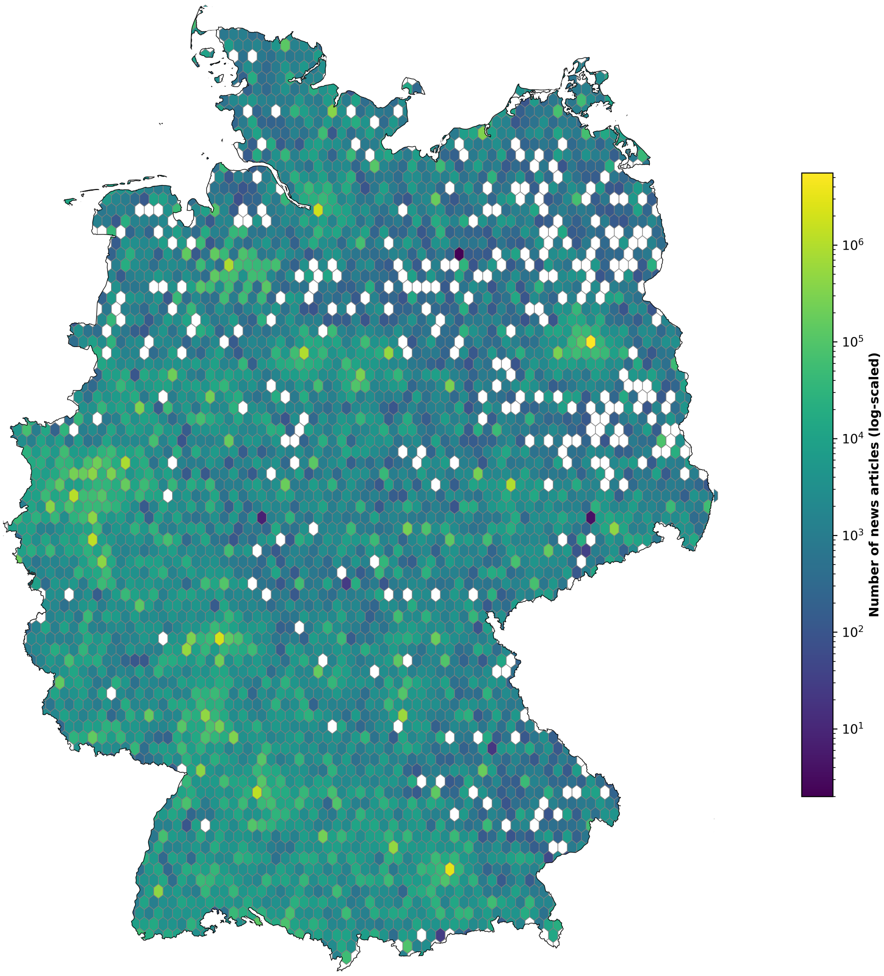

Figure 6: Number of news articles per location

In addition to mapping news articles per location, the dataset also uncovers relational patterns between places. To uncover the relationships and associations between different cities as represented in news coverage, we constructed a network of co-occurrences of city names within the same news articles. By analyzing these co-occurrence patterns, we can identify how cities are linked through shared events, themes or narratives in the news. This network analysis helps reveal underlying geographic and social dynamics, highlighting regions that frequently appear together in news texts. It also provides insights into how news coverage reflects or influences perceptions of city interconnections. In total we found 62,440,126 connections between locations in 25,870,054 news articles. Figure 7 illustrates this network, showing strong connections between major German metropolises. The visualization of links on the map is generated using a force-directed edge bundling algorithm. Notably, the strongest connections exist between Berlin and Munich, with 762,412 co-occurrences; Hamburg and Berlin, with 549,014 co-occurrences; Frankfurt and Berlin, with 441,034 co-occurrences; and Frankfurt and Munich, with 413,452 co-occurrences. Additionally, the network reveals connections between larger cities and their surrounding areas, highlighting regional interdependencies.

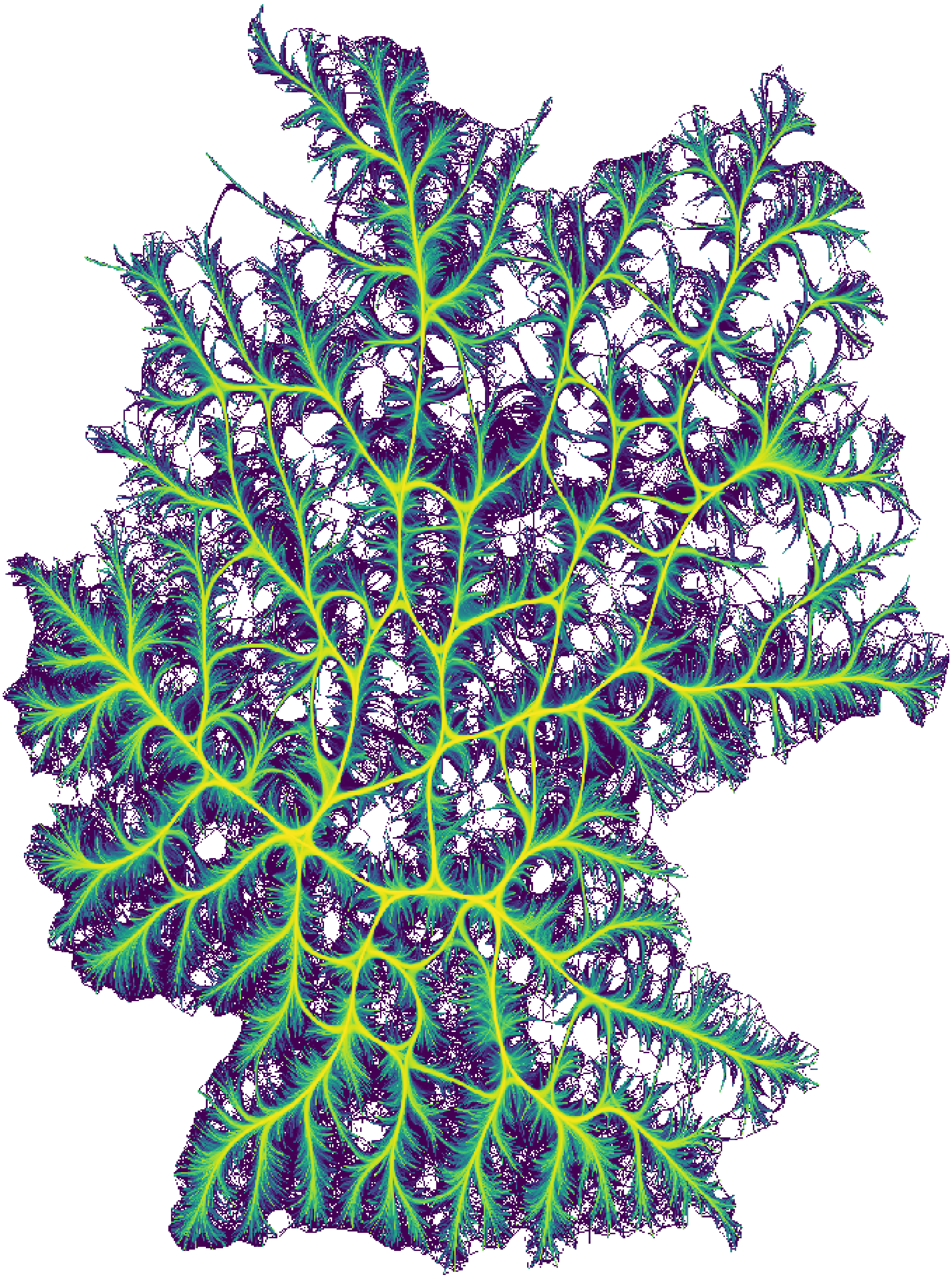

Figure 7: Network of co-occurrences of locations in articles

Data Records

The data is stored in a repository hosted by the University of Giessen[32]. The dataset consists of an SQLite database and a Usearch vector database, which together provide comprehensive data storage and semantic search capabilities. The SQLite database contains structured information about articles and their associated geographic locations, while the vector database enables efficient semantic search through vector representations of the articles. The article titles, texts and excerpts associated with this data can be retrieved directly from Common Crawl and linked to this dataset using the provided IDs.

SQLite Database

The Articles table contains information about all articles, including their unique identifiers, URLs, and metadata. The schema of the table is as follows:

Column name	Description
id	Unique identifier for each article (UUID format).
url	Original URL of the article.
tags	Tags associated with the article.
categories	Categories associated with the article.
hostname	Hostname of the article's source.
date	Publication date of the article in ISO format (YYYY-MM-DD).
date_crawled	Date when the article was crawled in ISO format (YYYY-MM-DD HH:MM).

Table 2: Schema overview of Articles table

The Locations table contains information about geographic locations, including normalized names and geographic coordinates. The schema of the table is as follows:

Column name	Description
location_id	Unique identifier for each location.
loc_normal	Normalized name of the location (for geocoding purpose).
latitude	Latitude coordinate of the location.
longitude	Longitude coordinate of the location.
NUTS	Nomenclature of Territorial Units for Statistics identifier.
GEN	General Name of the NUTS location
ARS	Regional identification number

Table 3: Schema overview of Locations table

The Article_Locations table serves as a join table, linking articles from the Articles table to geographic locations in the Locations table. This table supports a many-to-many relationship between articles and locations. The schema of the table is as follows:

Column name	Description
article_id	Unique identifier of the article (foreign key to Articles.id).
location_id	Unique identifier of the location (foreign key to Locations.location_id).

Table 4: Schema overview of Article_Locations table

The Article_Vectors table serves as a bridge between the vector store and the SQLite database. In the vector store, the IDs correspond to hashed article_id values. The table schema is outlined in Table 5.

Column name	Description
article_id	Unique identifier of the article (foreign key to Articles.id).
hashed_id	Hashed version of the article_id used in the vector store

Table 5: Schema overview of Article_Vectors table

Usearch Vector Database

The Usearch[32] vector database enhances the dataset by enabling semantic search through vector representations of the articles. Each article in the SQLite database is associated with a high-dimensional vector in the vector database, capturing the semantic content of the article. This structure facilitates efficient similarity-based retrieval. To locate articles related to a specific topic or keyword, the keyword is transformed into a numerical query vector, which is then compared against the stored article vectors to measure similarity and identify relevant results. Figure 8 depicts the architecture of the vector database.

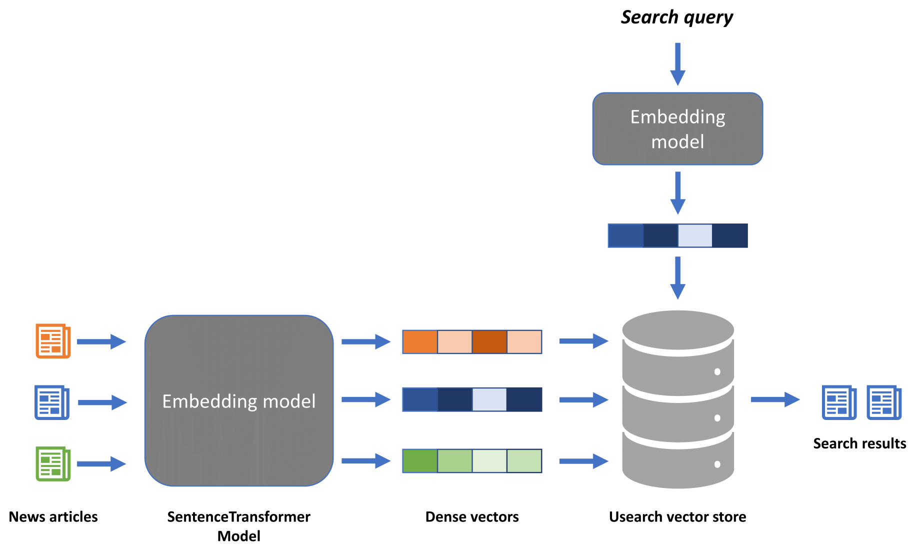

Figure 8: Vector database architecture (visual based on [33])

We provide example code for querying and filtering the database semantically as well as model recommendations. Moreover, we publish the vector data in different levels of granularity to meet diverse performance and storage requirements. Table 6 details the different versions of the vector database, leveraging quantization techniques to optimize for size, speed, and accuracy trade-offs.

Name	Quantization	Size	Distance metric
NewsIndex_f32	Float32	215 GB	Cosine distance
NewsIndex_int8	Int8	60GB	Inner product
NewsIndex_binary	Binary	14 GB	Hamming distance

Table 6: Overview of vector database quantizations

Embedding quantization is a method to reduce storage requirements and computational costs while maintaining sufficient accuracy for semantic search. Quantization techniques, such as reducing embeddings from 32-bit floating-point precision (Float32) to 8-bit integers (Int8) or binary representations, can reduce the size of the vector data significantly with some trade-off in retrieval precision.[34]

The integration of the vector database with the SQLite database enables advanced query capabilities. Users can perform semantic searches to find articles with similar content and then retrieve associated metadata and geographic information from the SQLite database. This combination facilitates a wide range of analyses, including geographic trends in article topics and content similarity analysis.

Technical Validation

To ensure the robustness and reliability of our dataset and methods, we conducted a series of validations targeting key components of the data processing pipeline. These validations assess the quality and consistency of the named entity recognition, geocoding, and vector database functionality, ensuring they meet the standards required for subsequent analyses**.** The validation processes were designed to evaluate the spatial, semantic, and temporal accuracy of our approach, emphasizing the dataset's ability to reflect real-world patterns and trends.

Named entity recognition and geocoding

In validating the named entity recognition (NER) and geocoding processes, we employed a log-log linear regression model. Figure 9 shows the relationship between number of articles and population size at NUTS-3 level. The analysis produced a coefficient of 0.99 ($p < 0.001$) for the relationship between population size (2022) and the pooled number of news articles, with an R^2 value of 0.507. This indicates that 50.7 % of the variance in news article counts is explained by population size, confirming that our geocoding process effectively captures the association between population and news coverage at NUTS-3 level. The significant coefficient of 0.99 suggests that the number of news articles scales nearly proportionally with population size. The consistent linear relationship serves as evidence that our geocoding method is capable of correctly attributing news articles to the relevant locations, confirming that it is neither overestimating nor underestimating the distribution of news coverage relative to population. This ensures that our geocoding approach is reliable for subsequent analyses that depend on accurate spatial allocation of news data.

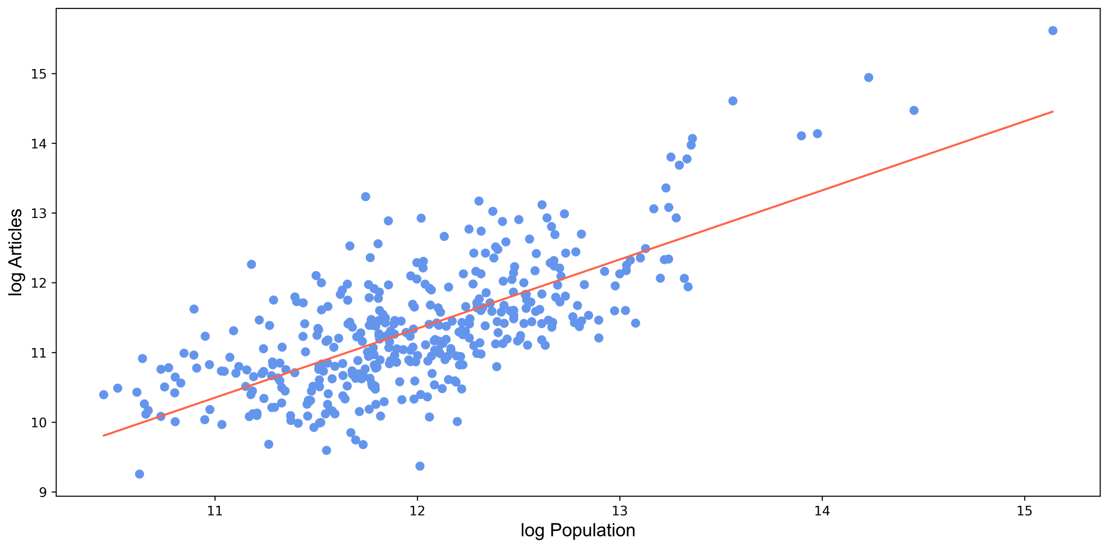

Figure 9: Number of articles and population size at NUTS-3 level

To validate the performance of our NER model, we conducted an evaluation using a randomly selected sample of 500 news articles, which were manually annotated for comparison. This manual annotation served as the ground truth against which the model's performance was assessed. The evaluation yielded an F1-score of 93.87%, indicating a high level of accuracy in recognizing entities. The balance between recall and precision suggests that the model performs consistently in identifying entities without significant bias toward either false positives or false negatives. Results of the evaluation are detailed in Table 7, showcasing the precision, recall, and F1-score metrics that confirm the model performance.

Precision	Recall	F1-Score
0.9363	0.9412	0.9387

Table 7: Performance metrics of NER model

Vector database

To validate the functionality and effectiveness of the vector database, we conduct a use case focusing on "heat pumps." A heat pump is a device that moves heat to warm or cool a building by extracting heat from the outside air or the ground and transferring it indoors. Heat pumps have garnered significant media attention in Germany in recent years, making them an informative and suitable example for this use case. Our aim is to demonstrate the precision of the semantic search capabilities and compare the extracted news volume against an independent data source—Google Trends—to ensure the robustness of the results. Using the vector database in 32-bit floating-point precision, we perform a semantic search for the term "heat pumps" in our German news dataset. By leveraging the pre-computed embeddings, the search algorithm retrieves articles whose semantic content closely aligns with the term, not just exact keyword matches. This allows us to capture articles that discuss heat pumps in various contexts, even when alternative phrasing or technical jargon is used.
We employ a two-step filtering process using the vector database to retrieve relevant articles. First, we use precomputed embeddings (Bi-encoder) to filter articles with a similarity score greater than 0.7. This initial step quickly narrows down the dataset by identifying articles semantically similar to "heat pumps" across various contexts, even when different terminologies are used.
Next, the remaining articles are fed into a more fine-grained reranker model.[35] This model performs a more precise evaluation, and we retain all articles with a similarity score greater than 0.1. This two-stage approach allows for a balance between computational efficiency and retrieval accuracy, ensuring that only relevant articles are selected for further analysis.

To evaluate the accuracy of different vector database configurations, we compare three precision levels: Binary, Int8, and Float32. We conduct an identical search across all three databases for the query "heat pumps", retrieving the 300,000 nearest results in each case. A threshold of k = 300,000 corresponds to a cosine similarity of 0.7 in the Float32 database. The retrieved results from each database are subsequently processed through the same reranker model. Table 8 shows a comparison of retrieval accuracy and database size across the three precision levels, with the Float32 configuration serving as the benchmark.

Precision	Retrieval accuracy	Size
Float32	100 %	215 GB
Int8	93.55 %	60 GB
Binary	51.14 %	14 GB

Table 8: Comparison of retrieval accuracy across different precision levels

Our results demonstrate that the Int8 precision offers an interesting balance between size and accuracy, achieving a performance retention of 93.55% while offering a 4x reduction in size compared to the Float32 configuration.

To externally validate the results, we compare the temporal distribution of heat pump-related news articles from our dataset against data from the Google Trends News Index for the same term in Germany over the corresponding time period. Google Trends provides a normalized measure of search interest, allowing us to benchmark the frequency of media coverage against public interest. Similarly, we normalized the Common Crawl news data by scaling article counts into percentages relative to the highest observed count, facilitating a comparative analysis of media coverage trends.

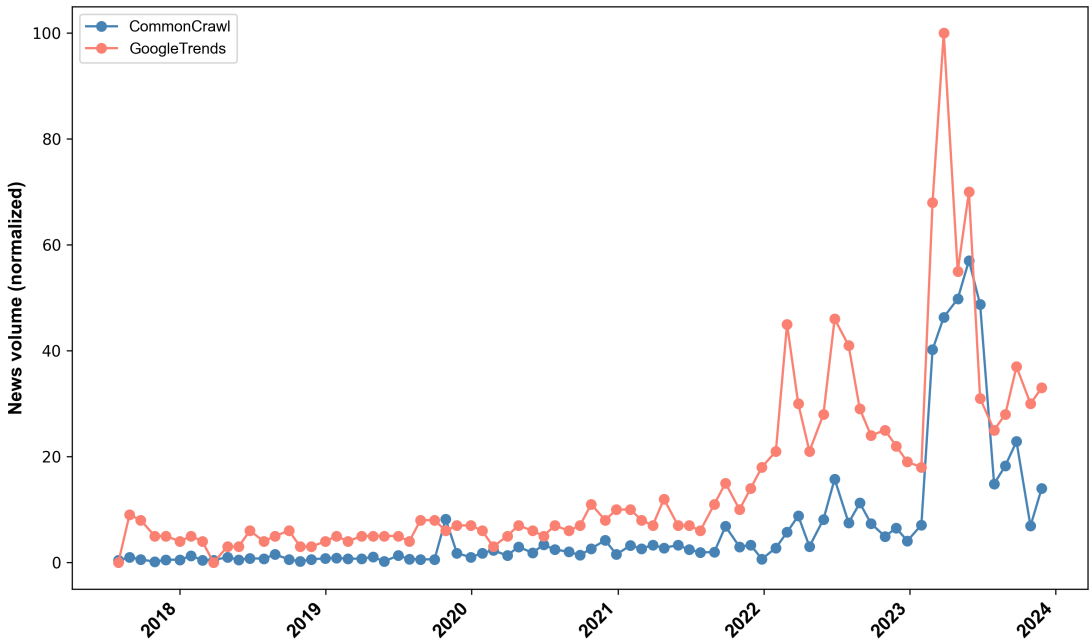

Figure 10: Comparison of news volumes

Figure 10 illustrates the comparison between the volume of "heat pumps" articles and Google Trends data from 2018 to 2024. The two datasets show a strong correlation, with peaks in media coverage closely mirroring peaks in Google search interest. This alignment is particularly evident in 2023, when the energy crisis and governmental debates surrounding the Buildings Energy Act significantly heightened public interest in renewable energy technologies.
The Pearson correlation coefficient between the monthly news article count and Google Trends data is 0.84, demonstrating a strong linear relationship. In sum, the use case highlights the reliability of the dataset and confirms the effectiveness of the filtering process.

Usage Notes

The dataset can be linked to other spatial data through the geographic information provided, such as the NUTS identifier. The dataset contains a large compilation of news articles metadata, many of which may not be relevant to every use case. In addition to utilizing semantic search via the vector database to produce relevant subsets of the news data tailored to a specific use case, users may also consider filtering the database by other available variables. For example, users might subset by news provider, location, or tag, among other possibilities. This will facilitate data handling. Since the dataset is based on Common Crawl News data, it does not include news articles behind paywalls. Article titles and texts can be retrieved directly from Common Crawl and linked to this dataset using the provided IDs.

Code Availability

All Python code produced for this project can be accessed on: https://github.com/LukasKriesch/CommonCrawlNewsDataSet.

Author contributions

S.L. and L.K. designed the study, wrote and reviewed the paper. L.K. managed the data, conducted the analysis, and produced the final dataset.

Competing interests

No potential competing interest was reported by the authors.

References

1. Benoit, K. Text as Data: An Overview. *The SAGE Handbook of Research Methods in Political Science and International Relations* 461–497 (2020) doi:10.4135/9781526486387.N29.
2. Grimmer, J., Roberts, M. E. & Stewart, B. M. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. (Princeton University Press, 2022).
3. Gentzkow, M., Kelly, B. & Taddy, M. Text as Data. *J Econ Lit* **57**, 535–74 (2019).
4. Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D. & Resch, B. The digital layer: alternative data for regional and innovation studies. *Spat Econ Anal* **18**, 507–529 (2023).
5. Kinne, J. & Lenz, D. Predicting innovative firms using web mining and deep learning. *PLoS One* **16**, e0249071 (2021).
6. Binsbergen, J. H. van, Bryzgalova, S., Mukhopadhyay, M. & Sharma, V. (Almost) 200 Years of News-Based Economic Sentiment. *NBER Working Papers* (2024).
7. Kriesch, L. & Losacker, S. Bioeconomy firms and where to find them. *REGION* **11**, 55–78 (2024).
8. Dahlke, J. *et al.* Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption. *Res Policy* **53**, 104917 (2024).
9. Blazquez, D. & Domenech, J. Big Data sources and methods for social and economic analyses. *Technol Forecast Soc Change* **130**, 99–113 (2018).
10. Abbasiharofteh, M. & Kriesch, L. *Not All Twins Are Identical: The Digital Layer of 'Twin' Transition Market Applications*. (2024).
11. McCombs, M. E. & Valenzuela, S. *Setting the Agenda: The News Media and Public Opinion*. (Polity Press, 2020).
12. Entman, R. M. How the Media Affect What People Think: An Information Processing Approach. *J Polit* **51**, 347–370 (1989).
13. McCombs, M. E. & Shaw, D. L. The agenda-setting function of mass media. *Public Opin Q* **36**, 176–187 (1972).
14. Baker, S. R., Bloom, N. & Davis, S. J. Measuring Economic Policy Uncertainty. *Q J Econ* **131**, 1593–1636 (2016).
15. Soo, C. K. Quantifying Sentiment with News Media across Local Housing Markets. *Rev Financ Stud* **31**, 3689–3719 (2018).
16. Lenz, D. & Winker, P. Measuring the diffusion of innovations with paragraph vector topic models. *PLoS One* **15**, e0226685 (2020).

17. Damstra, A. & Boukes, M. The Economy, the News, and the Public: A Longitudinal Study of the Impact of Economic News on Economic Evaluations and Expectations. *Communic Res* **48**, 26–50 (2021).
18. Ban, P., Fouirnaies, A., Hall, A. B. & Snyder, J. M. How Newspapers Reveal Political Power. *Political Sci Res Methods* **7**, 661–678 (2019).
19. Mueller, H., Rauh, C. & Seimon, B. Introducing a global dataset on conflict forecasts and news topics. *Data Policy* **6**, e17 (2024).
20. Ozgun, B. & Broekel, T. Assessing press releases as a data source for spatial research. *REGION* **9**, 25–44 (2022).
21. Ozgun, B. & Broekel, T. The geography of innovation and technology news - An empirical study of the German news media. *Technol Forecast Soc Change* **167**, 120692 (2021).
22. Berle, E. C. & Broekel, T. Spinning stories: Wind turbines and local narrative landscapes in Germany. *Technol Forecast Soc Change* **211**, 123892 (2025).
23. Ozgun, B. & Broekel, T. Saved by the news? COVID-19 in German news and its relationship with regional mobility behaviour. *Reg Stud* **58**, 365–380 (2024).
24. Rae, J. W. *et al.* Scaling Language Models: Methods, Analysis & Insights from Training Gopher. (2021).
25. Barbaresi, A. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the System Demonstrations* 122–131 (2021) doi:10.18653/V1/2021.ACL-DEMO.15.
26. Penedo, G. *et al. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale*. https://arxiv.org/abs/2406.17557v2 (2024).
27. Lopukhin, K. *Evaluating Quality of Article Body Extraction for Commercial Services and Open-Source Libraries*. (2019).
28. Raffel, C. *et al.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* **21**, 1–67 (2020).
29. Colvin, S. *et al.* Pydantic. https://docs.pydantic.dev/latest/ (2024).
30. Nominatim. Features. https://nominatim.org/ (2024).
31. Lee, S., Shakir, A., Koenig, D. & Lipp, J. Open Source Gets DE-licious: Mixedbread x deepset German/English Embeddings. https://www.mixedbread.ai/blog/deepset-mxbai-embed-de-large-v1 (2024).
32. Kriesch, L. & Losacker, S. A geolocated dataset of German news articles [Dataset]. https://doi.org/10.22029/jlupub-19573
33. Vardanian, A. USearch by Unum Cloud. Preprint at https://doi.org/10.5281/zenodo.7949416 (2023).
34. Allahyari, M. Semantic Document Search with ModernBERT and Milvus. https://github.com/mallahyari/modernbert-semantic-search?tab=readme-ov-file (2024).
35. Shakir, A., Aarsen, T. & Lee, S. Binary and Scalar Embedding Quantization for Significantly Faster & Cheaper Retrieval. *Hugging Face Blog* https://huggingface.co/blog/embedding-quantization (2024).
36. Laurer, M., van Atteveldt, W., Casas, A. & Welbers, K. Building Efficient Universal Classifiers with Natural Language Inference. (2023).