A diverse Multilingual News Headlines Dataset from around the World

Felix Leeb and **Bernhard Schölkopf**
Max Planck Institute for Intelligent Systems
Tübingen, Germany
fleeb@tue.mpg.de

Abstract

BABEL BRIEFINGS is a novel dataset featuring 4.7 million news headlines from August 2020 to November 2021, across 30 languages and 54 locations worldwide with English translations of all articles included. Designed for natural language processing and media studies, it serves as a high-quality dataset for training or evaluating language models as well as offering a simple, accessible collection of articles, for example, to analyze global news coverage and cultural narratives. As a simple demonstration of the analyses facilitated by this dataset, we use a basic procedure using a TF-IDF weighted similarity metric to group articles into clusters about the same event. We then visualize the *event signatures* of the event showing articles of which languages appear over time, revealing intuitive features based on the proximity of the event and unexpectedness of the event. The dataset is available on Kaggle and HuggingFace with accompanying GitHub code.

1 Introduction

Analyzing news headlines can be an invaluable source of data for a wide variety of natural language processing tasks such as bias detection (Gangula et al., 2019), topic classification (Rana et al., 2014), or event tracking (Qian et al., 2019). Furthermore, news headlines can provide insights for sociologists and political scientists about how people think about and discuss current events.

The coverage and discussion of current events varies significantly across different media outlets worldwide, however, these distinctions may be difficult to integrate in data mining or machine learning systems due to the language barrier. There are relatively few datasets offering extensive, diverse, and multilingual content (Kreutzer et al., 2022). This is especially problematic for natural language processing tasks, which have been shown to exhibit language biases (Gallegos et al., 2023).

We seek to address these limitations with a new dataset called BABEL BRIEFINGS, which is an accessible dataset representing a wide variety of languages and cultures. BABEL BRIEFINGS provides daily headlines of articles from across the world, originally written in one of 30 languages from 54 locations around the world published between August 2020 and November 2021, for a total of about 4.7 million distinct articles. Consequently, our dataset offers a rich source for analyses of world events, cultural narratives, media framing, and more.

We make this dataset available on Kaggle and HuggingFace for easy and open access under the CC BY-NC-SA 4.0 license [1], as well as providing all code used to collect and process the data on GitHub.

1.1 Related Work

Many comparable datasets focus either on *depth*, i.e., tracking a small number (or even a single) outlet over some time, or *breadth* for comparative studies of specific events. Meanwhile, our dataset covers a broad set of outlets in different languages over more than one year for both comparative and longitudinal studies.

Some of the related datasets publicly available include the News Category Dataset (Misra, 2022), BBC News Archive (Greene and Cunningham, 2006), AG News (Zhang et al., 2015), CC News (Hamborg et al., 2017). However, all of these datasets are mostly or entirely limited to English headlines and/or outlets. Meanwhile, datasets used in projects like Mazumder et al. (2014) or Leskovec et al. (2009) focus on collecting many sources in over a relatively short timespan (see the appendix for a comparison table).

A more global source of news events is offered

[1] This license permits non-commercial use as long as the dataset is credited and variants are licensed under the same terms.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 647–652
June 16-21, 2024 ©2024 Association for Computational Linguistics

by the GDELT project (Leetaru and Schrodt, 2013), which collects reports from around the world in a variety of languages. However, the GDELT dataset focuses on tracking events, rather than the news coverage thereof, making it more suitable for event forecasting rather than media coverage or training language models.

2 Dataset

2.1 Collection

The dataset was collected in three distinct steps. First, using the News API (News-API, 2023), we gathered the available headlines once a day for each combination of all 54 locations and each of the seven possible categories. Each API call returned a list of about 30-70 article headlines for a total of about 20k instances per day, usually featuring duplicate headlines across locations and categories.

Next, in a pre-processing step, duplicate occurrences of the same article were merged and listed in a list of `instances` (see below). The author names are anonymized (replacing the names with `author#[ID]` where the `ID` is identical for all articles with matching authors, but distinct otherwise).

Lastly, the final step involved the translation of non-English articles. Using Google Translate (Google, 2023), all articles not originally in English were translated for convenience. Notably, News API appears to only collect articles of a single language for each of the locations, making translation straightforward. Unfortunately, some of the language selections by News API do not seem to fully reflect the local news in a given location (for example Malaysia's articles are all in English), although the headline subjects appear curated for the assigned location.

2.2 Structure

BABEL BRIEFINGS is structured as a collection of 54 JSON files, one for each location. Each file contains a list of headlines of articles that first occurred in the corresponding location, each of which is represented as a JSON object with the following properties: `title`, `description`, `content`, `url`, `urlToImage`, `publishedAt`, `author`, `source`, `instances`, and `language`. For articles that are originally in a language other than English, the translated `title`, `description`, and `content` are also included as `en-title`, `en-description`, and `en-content` respectively.

Since each article may appear in multiple locations, categories, or over multiple days, the `instances` property lists the properties `location` and `category` for each instance when the article was collected with timestamp `collectedAt`. The `source` property is an object containing the `id` and `name` of the news source. The `language` property is the original language (in ISO 639-1 format) of the article, which is assigned automatically based on the location. A single category is assigned to each instance automatically by News API, and are one of: `business`, `entertainment`, `general`, `health`, `science`, `sports`, or `technology`.

Notably, the `content` (and `en-content`, when present) properties contain nonsense data for articles in languages that use a non-latin alphabet, such as Chinese or Arabic. This is due to a flaw in the News API processing. For details check out the dataset's readme.

2.3 Statistics

In total, we collected a total of 7,419,089 instances of 4,719,199 distinct articles between 8 August 2020 and 29 November 2021, with a breakdown by language in table 1. More detailed statistics are available in the dataset readme.

3 Analysis

A particularly interesting type of analysis enabled by this dataset is the longitudinal comparison of how the same news event is reported in different languages and around the world. To illustrate, let's consider a basic example. We begin by clustering individual articles that discuss the same event. Then, we analyze the distribution and frequency of articles from different countries over time, focusing on that specific event.

To cluster articles that are about the same event, we begin by extracting a bag of words from the article's (English) title where each word is lemmatized as well as removing punctuation and common stopwords (such as "the" or "a"). We use Term Frequency-Inverse Document Frequency (TF-IDF) (Salton et al., 1975) to define the relevance R_d of each token relative to the other tokens of the articles that occurred on the same day d.

$$R_d(w) = \frac{\mathrm{tf}(w,d)}{\sum_{d'} \mathrm{tf}(w,d')} \cdot \log \frac{N}{\mathrm{df}(w)} \qquad (1)$$

where $\mathrm{tf}(w,d)$ is the number of times the word w occurs in the day d, $\mathrm{df}(w)$ is the number of days

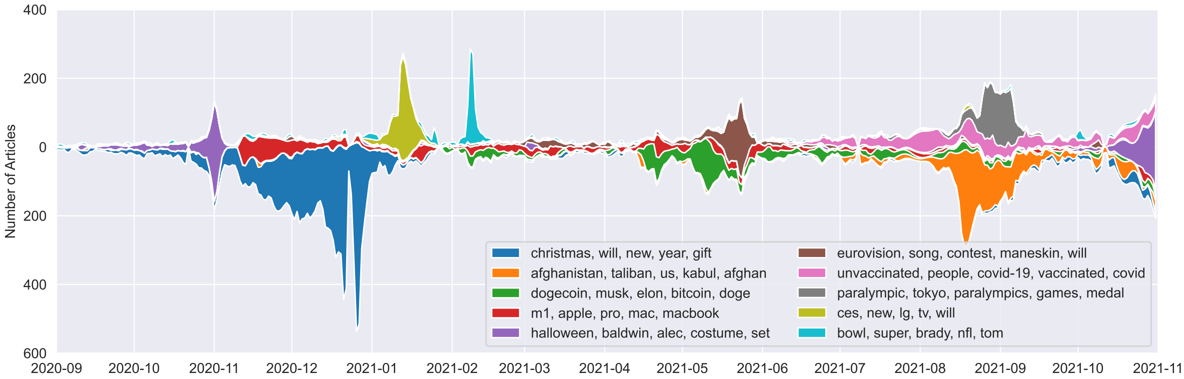

Figure 1: Streamplot showing how many articles appear for some of the most popular events in the dataset, when clustering articles by their titles, with most common tokens for each cluster shown in the legend. Note the qualitative similarity between the news coverage over time of these events and the memes of Leskovec et al. (2009), demonstrating the potential of this dataset for studying the evolution of major news coverage over time across the world.

Language	**Articles**
English	1,128,233
Spanish	455,952
French	288,328
Chinese	270,887
German	259,718
Portuguese	243,829
Arabic	178,854
Indonesian	131,252
Italian	129,005
Turkish	122,724
Greek	119,940
Japanese	118,475
Polish	116,904
Russian	113,395
Dutch	104,031
Thai	90,708
Swedish	86,838
Korean	83,090
Serbian	80,040
Hungarian	73,509
Czech	70,647
Hebrew	67,794
Bulgarian	67,223
Ukrainian	65,610
Romanian	54,601
Norwegian	46,804
Slovak	43,057
Latvian	40,006
Lithuanian	34,719
Slovenian	33,026

Table 1: Number of articles by language.

in which the word w occurs, and N is the total number of days in the dataset.

Using the TF-IDF scores for each word, we define a relevance score $\hat{R}_d(x) = \sum_{w \in x} R_d(w)$ for an article x that occurs first on day d as the sum of the TF-IDF scores of the words in its title. Furthermore, we define a similarity criterion between two articles as the ratio between the sum of all words that occur in both articles weighted by the relevance of each word and the largest relevance score between the two articles.

$$\mathrm{sim}(x, x') = \frac{\sum_{w \in x \cap x'} R_d(w)}{\max(\hat{R}_d(x), \hat{R}_d(x'))} \qquad (2)$$

If this ratio is greater than some threshold ($= 0.25$ in our experiments), we consider the two articles to be in the same group. This means that if the candidate articles have significant overlap between words weighted by how specific those words are to the day. For the top ten articles with the highest relevance scores every day, we identify all articles in the dataset which, based on our similarity criterion are in the same group to form an event cluster.

Figure 1 presents clusters of such articles, identifying the top TF-IDF scores where the clusters are largest—that is, events with the most articles published about them.

Next, we take a closer look at a few of the largest clusters in figures 2-5. We visualize the *event signatures*, which show how the coverage of the same event varies across different languages by how many articles are published every day. For each of the four examples, the plot shows a streamplot breaking down how many articles were published

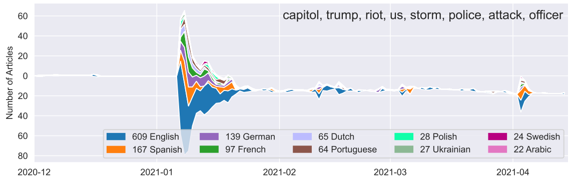

Figure 2: Articles reporting on riots in Washington DC on 6 January 2021. Note how the event is reported in many different languages, but the majority of articles are in English. Additionally, there are several subsequent smaller spikes corresponding to related events, such as the beginning of the formal investigation into the riots.

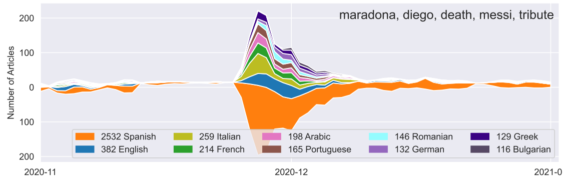

Figure 3: Articles reporting on Diego Maradona's death on 25 November 2020 (and his declining health in the weeks before). Note how in after a few weeks only Spanish articles about the topic continue to appear, underscoring the relative importance of the event in Spanish-speaking countries.

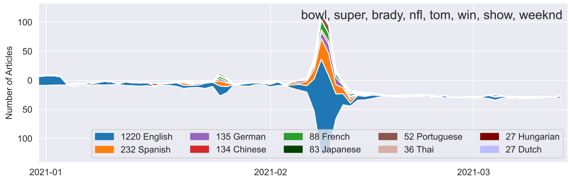

Figure 4: Articles reporting on the Super Bowl on 7 February 2021. Note how unlike unexpected events (such as in figure 2), there is a considerable lead up to the event before the peak, showing the media's anticipation of the event.

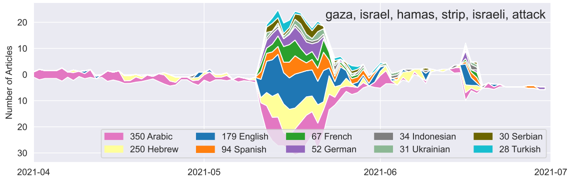

Figure 5: Articles reporting on a crisis between Israel and Gaza on 10 May 2021. Note the prolonged spike for the duration of the crisis, as well as the significant number of articles in Arabic and Hebrew.

for each of the top ten most common languages for the event as well as the most frequent tokens occurring that cluster in the top right. One interesting result from this precursory analysis is a distinct qualitative difference in the event signatures of "expected" events (such as in figure 4) compared to "unexpected" events (such as in figures 2 and 5). For expected events, there is a clear lead-up to the event, with a peak on the day of the event, and a sharp drop-off afterwards. Meanwhile, unexpected events show a sudden spike in coverage, followed by a gradual decline over time. This provides a demonstration of the types of analyses that can be conducted with this dataset, offering insights into the diversity and scope of global news coverage.

4 Conclusion

In this paper, we introduce a dataset of news headlines from around the world called BABEL BRIEFINGS. The dataset can readily be used for a wide variety of both supervised and unsupervised natural language processing tasks. For example, the included category, location, and language labels can directly be used for article categorization, location classification, or language detection. However, the dataset also enables more nuanced analyses of global news coverage, such as tracking the evolution of events over time, comparing the coverage of events across different countries and languages, or identifying cultural biases in reporting.

Despite the breadth across languages and time, our dataset is limited to the headlines and short descriptions of news articles. However, URLs to the full articles are included, and since many outlets are incentivized to make their headlines as informative as possible, headlines alone are already a rich source of information for many purposes. Additionally, the dataset is limited to 54 locations, which is a significant improvement over existing datasets which are often limited to a single country or outlet. There are some minor issues with the News API, for example that for each location only a single language is represented. We aim to mitigate this issue by collecting headlines directly from the RSS feeds of individual outlets from around the world. However, this may come at the cost of consistency across sources around the world.

In summary, our dataset is a powerful tool for studying the nuances of global news coverage when breaking beyond the language barrier. It provides a simple yet rich foundation for capturing cultural differences in news reporting, offering invaluable data and insights for researchers in the fields of natural language processing, as well as social sciences like media studies or international relations.

References

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy. Association for Computational Linguistics.

Google. 2023. Google translate. https://translate.google.com. Accessed: 30 October 2023.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384.

Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. 2017. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506.

Sahisnu Mazumder, Bazir Bishnoi, and Dhaval Patel. 2014. News headlines: What they can tell us? In *Proceedings of the 6th IBM Collaborative Academia Research Exchange Conference (I-CARE) on I-CARE 2014*, pages 1–4.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

News-API. 2023. News api. https://newsapi.org/. Accessed: 8 August 2020.

Yu Qian, Xiongwen Deng, Qiongwei Ye, Baojun Ma, and Hua Yuan. 2019. On detecting business event from the headlines and leads of massive online news articles. *Information Processing & Management*, 56(6):102086.

Mazhar Iqbal Rana, Shehzad Khalid, and Muhammad Usman Akbar. 2014. News classification based on their headlines: A review. In *17th IEEE International Multi Topic Conference 2014*, pages 211–216. IEEE.

Gerard Salton, A Wong, and C S Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

A News Headline Dataset Comparison

Dataset Name	Size	Sources	Language/s	Time Span
BABEL BRIEFINGS (ours)	4.7M	Worldwide	30 languages	Aug 2020 - Nov 2021
News Category Dataset (Misra, 2022)	210k	HuffPost	English only	2012-2022
BBC News Archive (Greene and Cunningham, 2006)	2225	BBC	English only	2004-2005
AG News (Zhang et al., 2015)	128k	>2000	English only	2004
CC News (Hamborg et al., 2017)	708k	Worldwide	English only	Jan 2017 - Dec 2019
Mazumder et al. (2014) Dataset	1.5M	87 Indian sources	English only	Jan - Jun 2014
Leskovec et al. (2009) Dataset	90M	US news + blog sites	English only	Aug - Oct 2008
GDELT Project (Leetaru and Schrodt, 2013)	>326M	Worldwide	>100 Languages	since 1979

Table 2: Comparison of various existing datasets similar to BABEL BRIEFINGS