LLNewsBias: A Multilingual News Dataset for Lifelong Learning

Swati Swati
swati.swati@unibw.de
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute and
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

The rise of digital media enhances information accessibility but also introduces challenges related to the quality and impartiality of news reporting, particularly regarding biases that influence public perception during key global events. In response, this study introduces *LLNewsBias*, a dataset designed to detect and analyze political bias in multilingual news headlines, covering four major events from 2019 to 2022 — Brexit, COVID-19, the 2020 U.S. election, and the Ukraine-Russia war. With over 350,000 headlines in 17 languages, annotated with bias labels, this dataset is compiled using Media Bias/Fact Check and Event Registry. Our contributions include a structured framework for data collection and organization, enabling event-wise and year-wise analysis while supporting lifelong learning. We also highlight potential use cases that demonstrate the dataset's utility in advancing bias prediction models, multilingual adaptation, and model robustness. Additionally, we discuss the dataset's limitations, addressing potential biases, sample size constraints, and contextual factors. This work provides a valuable resource for improving bias detection in dynamic, multilingual news environments, contributing to the development of more accurate and adaptable models in natural language processing and media studies. For code and additional insights, visit: https://github.com/Swati17293/LLNewsBias

Keywords

Dataset, News, Bias, Multilingual, Headline, Low-resource, Media Bias, News Bias, Continual Learning, Lifelong Learning

1 Introduction

The rapid growth of digital media has greatly enhanced the accessibility of information, but it has also introduced significant challenges concerning the quality and impartiality of news reporting. Political bias in news content is particularly concerning, as it has the potential to influence public perception and shape societal narratives, especially around key global events. Understanding and predicting such biases, particularly in multilingual contexts where biases can manifest differently across cultural and linguistic boundaries, is essential for promoting fair and balanced journalism. Traditional approaches to bias detection often rely on monolingual datasets and static models that may not effectively capture the evolving nature of news content [6]. These limitations underscore the need for more robust datasets and methodologies that can adapt to the dynamic and multilingual landscape of modern news reporting.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.sikdd.8

In this study, we address these challenges by introducing a novel dataset *LLNewsBias* specifically designed for the detection and analysis of political bias in multilingual news headlines. Our dataset spans four major global events from 2019 to 2022: Brexit, COVID-19, the 2020 U.S. election, and the Ukraine-Russia war, capturing a wide range of political discourse across 17 languages. To collect this dataset, we use Media Bias/Fact Check for the assignment of bias labels, and Event Registry [2] for the extraction of relevant headlines and metadata. The resulting dataset is not only comprehensive in its linguistic diversity but also structured to support both event-wise and year-wise analyses, with an emphasis on lifelong learning.

1.1 Contributions

Our study makes the following contributions:

- **Multilingual bias-annotated dataset:** We introduce a multilingual bias-annotated dataset containing over 350,000 news headlines in 17 languages, each annotated with political bias labels.
- **Data collection and organization framework:** We present a structured framework for data collection and organization, enabling event-wise and year-wise analysis while ensuring adaptability for lifelong learning.
- **Potential use-cases:** We outline several potential applications of our dataset, highlight its potential for advancing lifelong learning models, particularly in bias prediction, multilingual adaptation, and model robustness.
- **Discussion of limitations:** We identify and discuss the dataset's limitations, such as biases in data collection, sample size constraints, and contextual influences, offering a transparent assessment of its applicability.

In summary, our paper introduces a comprehensive dataset and a framework for the study of political bias in multilingual news headlines. By focusing on key global events and providing support for lifelong learning, our study contributes to the ongoing effort to develop more accurate and adaptable models for bias detection in diverse linguistic and cultural contexts.

2 Related Work

Several datasets focus on news articles and political bias [5], but there is a notable scarcity of multilingual, bias-annotated datasets designed for lifelong learning [4]. While resources like the *media bias chart* by Ad Fontes Media and PolitiFact provide insights into bias, they are often limited to English-language sources or specific fact-checked claims, lacking the continuous, event-centric data necessary for broader analysis. GDELT [3], a large-scale event-oriented news dataset, covers multiple languages but focuses on location, network, and temporal attributes rather than political bias or the event-outlet relationship. Existing multilingual datasets are often domain-specific [1], limiting their utility for general bias analysis. In contrast, LLNewsBias dataset fills these gaps by offering a generalized, multilingual, and bias-annotated data designed for event-wise and year-wise analyses, particularly suited for lifelong learning models.

3 Dataset Description

In this section, we introduce our dataset *LLNewsBias* and describe the framework used for its collection and organization. We begin by detailing the primary data sources that form the foundation of this dataset. Following this, we present a comprehensive overview of the data collection process, with a focus on the methodologies employed to ensure robustness and reliability. Finally, we provide an in-depth overview of the dataset's structure, including its directory organization, file contents, and the various ordering methods applied to facilitate detailed analysis. Our dataset is documented in accordance with the FAIR Data Principles.

3.1 Primary Data Sources

In this section, we outline the two primary data sources used in our study: Media Bias/Fact Check (MBFC) and Event Registry (ER). MBFC serves as the bias rating portal, providing bias labels for selected media outlets, while ER is used to extract the headlines and corresponding metadata from articles published by these outlets.

3.1.1 ***Media Bias/Fact Check****.* For bias labeling in this study, we utilized Media Bias/Fact Check (MBFC), a well-established platform known for its comprehensive coverage and frequent updates. Although other platforms like allsides.com and adfontesmedia.com also provide bias ratings, MBFC was selected for its reliability and particular focus on low-resource languages. MBFC assigns bias labels based on political orientation and evaluates outlets for credibility and factual accuracy. These labels are determined by a team of contractors and volunteers who follow a standardized methodology, ensuring that the ratings are both consistent and dependable for our analysis.

3.1.2 ***Event Registry****.* In this study, we use Event Registry [2] platform as the primary source for collecting multilingual news headlines. It aggregates content from over 150,000 news sources across more than 60 languages, making it an ideal resource for analyzing bias in diverse and low-resource languages. Apart from the headlines, it allows access to numerous metadata such as publication date, news category, and political bias. By leveraging its Python API, we efficiently filtered and extracted headlines relevant to our study. This ensured a comprehensive dataset that supports the analysis of bias in a lifelong learning setup, exploring how emerging events and domain shifts influence the performance of bias prediction models over time.

3.2 Data Collection Framework

Our data collection framework as depicted in Figure 1, is designed to support both event-wise and year-wise analyses, with the additional capability of facilitating lifelong learning.

For data collection, we begin by defining two sets: a set of significant global events ($E = \{e_1, e_2, ..., e_n\}$), and a set of years ($Y = \{y_1, y_2, ..., y_m\}$), where n and m represent the total number of events and years, respectively. We then use the Media Bias/Fact Check (MBFC) platform to select media outlets ($O = \{o_1, o_2, ..., o_p\}$) and determine their respective political bias, with p as the total number of outlets. To maintain data reliability, we exclude outlets labeled as questionable and assign each remaining outlet $o_i \in O$ a bias label $b_i \in B$, where $B = \{b_1, b_2, ..., b_q\}$ represents the set of bias labels, with q representing the number of distinct bias labels.

Next, we define a temporal query Q_t to extract article headlines ($H = \{h_1, h_2, ..., h_r\}$), where r represents the total number of headlines retrieved from the Event Registry (ER). The query Q_t is formulated as:

$$Q_t = \{Q_e, Q_o, Q_{cat}, Q_{dt}\} \tag{1}$$

where Q_e, Q_o, Q_{cat} specify the event, media outlet, and news categories (limited to those classified as 'news' by ER $Q_{cat} = \{\text{`}politics\text{'}, \text{`}business\text{'}, \text{`}sports\text{'}, \text{`}arts\ and\ entertainment\text{'}, \text{`}science\text{'}, \text{`}technology\text{'}, \text{`}health\text{'}, \text{`}environment\text{'}\}$), respectively. The time constraint is represented as $Q_{dt} = [Q_{sd}, Q_{ed}]$, where Q_{sd} and Q_{ed} denote the start and end dates. To scrape all the article headlines (H), we utilize Q_t to query ER.

We then associate the extracted headlines H with the corresponding bias labels in B and structure the dataset according to two classification types: event-wise and year-wise. To organize the data, we define an event-based order O_{event} and a year-based order O_{year} as follows:

$$O_{event} = \{e_1 \to e_2 \to ... \to e_n\} \tag{2}$$

$$O_{year} = \{y_1 \to y_2 \to ... \to y_m\} \tag{3}$$

For lifelong learning, we design the dataset to be extendable, allowing for the integration of new events and years as they emerge, denoted by $E' \subseteq E$ and $Y' \subseteq Y$, where E' and Y' represent the sets of newly added events and years.

We designed the dataset with a flexible framework that allows for the seamless integration of new events and years as they emerge, represented as $E' \subseteq E$ and $Y' \subseteq Y$, where E' and Y' denote the newly added events and years. This structured approach ensures scalability for continuous learning without requiring major restructuring and supports the training of adaptive models capable of integrating new information effectively. Unlike standard multi-year datasets, our dataset includes annotations that facilitate contextual understanding, enabling models to learn from historical data while adapting to evolving trends and patterns in news reporting. This ensures that the models remain relevant as new information becomes available.

Finally, we split the dataset into training and test sets using a stratified sampling approach to ensure the preservation of bias label distributions across both events and years. We perform this step as it is critical for maintaining the integrity of the model training process in a lifelong learning context.

3.3 Data Synopsis and Structure

In this section, we present an overview of the data and explain how it is systematically organized, making it easier to understand both the content and format of our dataset.

3.3.1 ***Data Synopsis****.* The dataset features 356,060 headlines on four major events from 2019 to 2022: *Brexit*, *COVID-19*, *the election*, and *the Ukraine-Russia war*. These headlines, sourced from 45 unique news outlets in 17 different languages, are annotated with 3 political bias labels: *Left Centre*, *Least Biased*, and *Right Centre* covering diverse topics such as *politics*, *business*, *arts and entertainment*, *sports*, *science*, *technology*, *health*, and *environment*. The dataset is structured into 7 distinct columns within .csv files. Table 1 presents a comprehensive summary of the dataset statistics.

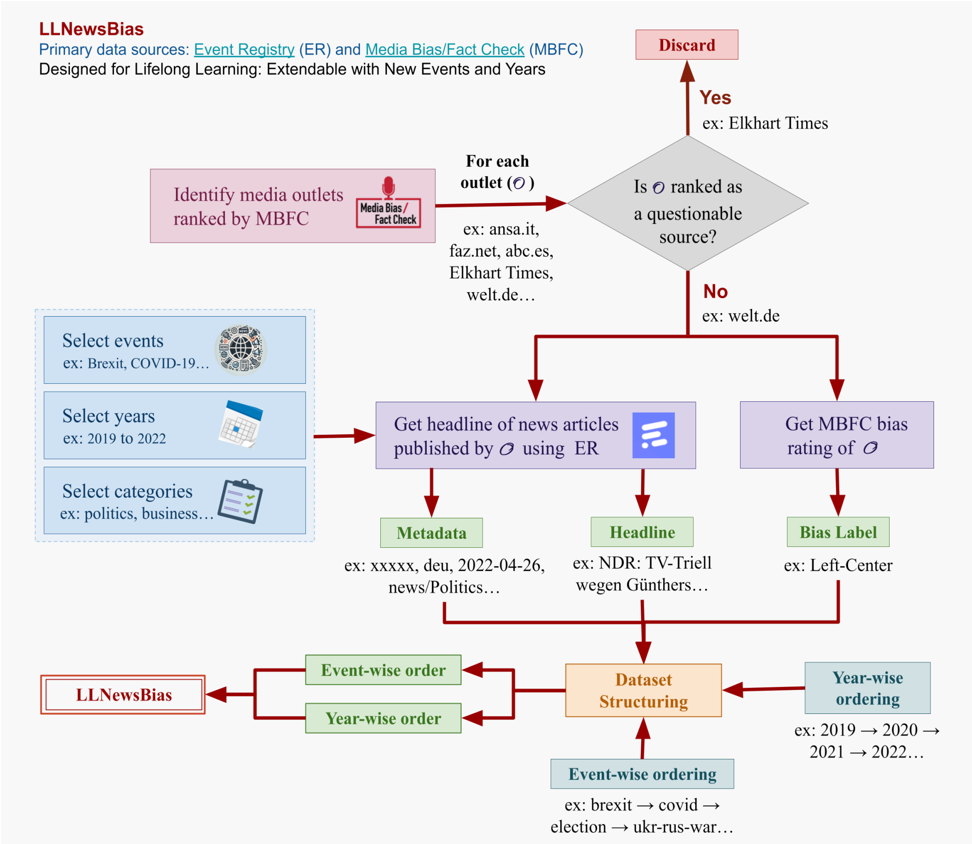

Figure 1: Data Collection Framework. The framework uses MBFC for bias labeling and ER for headline retrieval.

Table 1: Summary of Dataset Statistics.

Language-wise Distribution			
Catalan	882	Romanian	17,038
Croatian	13,929	Russian	10,511
Czech	1,876	Slovak	5,642
Danish	4,330	Spanish	83,940
Dutch	10,905	Swedish	6,441
Finnish	1,512	Ukrainian	10,616
French	85,007	Italian	48,450
Hungarian	105		
Event-wise Distribution			
Brexit	32,286	COVID	309,329
Election	3,829	Ukraine	10,616
Year-wise Distribution			
2019	20,664	2021	4,638
2020	258,871	2022	71,887

3.3.2 ***Directory Structure***. The dataset is organized in a main 'data' directory with subdirectories categorized by events ('brexit', 'covid', 'election', 'ukr-rus-war') and years (2019-2022). Additional subdirectories consolidate data across all events (ordered_events) and all years (ordered_years). Each subdirectory contains .csv files for training and testing, structured across the following columns.

- **news outlet**: The name of the news outlet.
- **article_ID**: A unique identifier for the raw news article in the Event Registry platform from which the headlines are extracted.
- **language**: The source language of the published news article.
- **date**: The date on which the news was published.
- **headline_text**: The text of the news headline.
- **news_category**: The category assigned by Event Registry.
- **political_bias**: The political bias of the news outlet as provided by the bias rating portal Media Bias/Fact Check.

The dataset is annotated with bias labels: Left Centre (LC), Least Biased (LB), and Right Centre (RC). To ensure model robustness across varying data distributions, we concatenate and shuffle files for each event and year in four distinct random orders. This prevents overfitting to specific sequences and helps evaluate generalization across diverse configurations. While chronological order is ideal for practical use, this randomized approach tests broader performance, with the original event and year splits provided for user flexibility.

Event-wise Ordering:

(1) brexit → covid → election → ukr-rus-war
(2) election → covid → ukr-rus-war → brexit
(3) brexit → ukr-rus-war → election → covid
(4) covid → brexit → ukr-rus-war → election

Year-wise Ordering:

(1) 2019 → 2020 → 2021 → 2022
(2) 2021 → 2020 → 2022 → 2019
(3) 2019 → 2022 → 2021 → 2020

(4) 2020 → 2019 → 2022 → 2021

The dataset captures the distribution of headlines related to various events over the years, reflecting the temporal dynamics of news coverage and the evolving reporting on these events. The differences in coverage levels reveal important patterns in media attention, which are essential for developing datasets that support lifelong learning models.

4 Potential Use-Cases

Our dataset introduced in this study has a wide range of potential use-cases, particularly in the fields of natural language processing and media studies. It is particularly valuable for research and applications that require understanding and predicting news bias in a continual, multilingual environment. Below we list some potential use cases:

- **Lifelong learning for news bias prediction:** Our dataset is ideal for developing and testing lifelong learning models. It allows models to adapt to new events and evolving entities. With its year-wise structure from 2019 to 2022, the dataset addresses the challenges of emerging events and domain shifts (e.g., Brexit, COVID-19, Ukraine-Russia War), providing the data needed to develop and evaluate robust models.

- **Domain Adaptation in Multilingual Contexts:** Our dataset enables researchers to investigate domain adaptation techniques in a multilingual context, featuring headlines in 17 languages. This facilitates the development of models that generalize across languages and adapt to various cultural and political contexts, ensuring accurate bias prediction. It addresses the challenges faced by generic models in the news domain, which often struggle with topic and language diversity.

- **Sparse Experience Replay for Continual Learning:** Our dataset is particularly well-suited for the news domain, supporting efficient experience replay by allowing the selection of specific topics and categories. With its event-wise and year-wise classifications, our dataset enhances memory utilization, improves generalization, reduces catastrophic forgetting, and ensures that models remain accurate and up-to-date in real-time applications.

In a nutshell, our dataset serves as a valuable resource for advancing news bias prediction, particularly in the context of lifelong learning, by providing a flexible framework for integrating new events and years. Unlike many news-based datasets with timestamps, it offers structured annotations and contextual information that enhance the understanding of evolving trends in news coverage, making it particularly suitable for lifelong learning applications. It supports a range of research activities, from model development and evaluation to the exploration of new techniques for handling dynamic and multilingual news environments.

5 Limitations

Several limitations are associated with the dataset presented in this article and should be carefully considered in any further research or analysis:

- **Data Collection Issues:** The dataset was gathered using Media Bias Fact/Check (MBFC) and the paid version of Event Registry (ER). MBFC is publicly accessible, while ER provided comprehensive but limited coverage, potentially missing relevant articles. The use of ER's paid version also restricted the extent of data collection.
- **Sample Size:** The dataset is constrained by its focus on four major events over a span of four years. This limited number of events and time frame may not fully capture the broader spectrum of news and media biases, affecting the diversity of the samples.
- **Biases:** Selection bias is a significant factor, as only news outlets labelled by MediaBiasFactCheck were included. This restriction may limit the number of languages and perspectives represented in the dataset, thereby influencing the overall analysis.
- **Contextual Factors:** The dataset is limited by its temporal scope, covering only four specific events over four years. While it reflects the dynamic nature of news media, it does not account for all future events and years to come.

6 Conclusions

In this study, we present LLNewsBias, a comprehensive dataset designed to tackle the challenges of detecting and analyzing political bias in multilingual news headlines. By spanning four major global events from 2019 to 2022 across 17 languages, this dataset provides a valuable resource for research in natural language processing and media studies. Our framework supports both event-wise and year-wise analysis, emphasizing lifelong learning and enabling models to adapt continuously to new data. The dataset's potential use cases include enhancing bias prediction models, facilitating domain adaptation in multilingual contexts, and improving model robustness. While LLNewsBias offers significant contributions, we also acknowledge limitations such as potential biases in data collection, sample size constraints, and contextual factors. Addressing these challenges in future work will be crucial for maximizing the dataset's impact, ultimately contributing to fairer and more balanced journalism.

7 Acknowledgments

This work was supported by the Slovenian Research Agency and National grants (CRP V2-2272; V5-2264; CRP V2-2146) and by the European Union through enrichMyData EU HORIZON-IA project under grant agreement No 101070284 and ELIAS HORIZON-RIA project under grant agreement No 101120237.

References

[1] Jason Armitage, Endri Kacupaj, Golsa Tahmasebzadeh, Swati, Maria Maleshkova, Ralph Ewerth, and Jens Lehmann. 2020. Mlm: a benchmark dataset for multitask learning with multiple languages and modalities. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2967–2974.

[2] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, 107–110.

[3] Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: global data on events, location, and tone, 1979–2012. In *ISA annual convention*. Vol. 2, 1–49.

[4] Swati Swati, Adrian Mladenić Grobelnik, Dunja Mladenić, and Marko Grobelnik. 2023. A commonsense-infused language-agnostic learning framework for enhancing prediction of political bias in multilingual news headlines. *Knowledge-Based Systems*, 277, 110838.

[5] Swati Swati, Dunja Mladenić, and Tomaž Erjavec. 2021. Eveout: an event-centric news dataset to analyze an outlet's event selection patterns. *Informatica*, 45, 7.

[6] Swati Swati, Dunja Mladenić, and Marko Grobelnik. 2023. An inferential commonsense-driven framework for predicting political bias in news headlines. *IEEE Access*.