SAFARI: Cross-lingual Bias and Factuality Detection in News Media and News Articles

Dilshod Azizov[1*], **Zain Muhammad Mujahid**[1*], **Hilal AlQuabeh**[1],
Preslav Nakov[1] **and Shangsong Liang**[1,2†]
[1]Mohamed bin Zayed University of Artificial Intelligence, UAE
[2] Sun Yat-sen University, China
{dilshod.azizov, zain.mujahid, hilal.alquabeh
and preslav.nakov}@mbzuai.ac.ae, liangshangsong@gmail.com

Abstract

In an era where information is quickly shared across many cultural and language contexts, the neutrality and integrity of news media are essential. Ensuring that the content of the media remains objective and factual is crucial to maintaining public trust. With this in mind, we introduce **SAFARI**(CroSs-lingual BiAs and Factuality Detection in News MediA and News ARtIcles), a novel corpus of news media and articles for predicting political bias and the factuality of the reporting in a cross-lingual setup. To our knowledge, this corpus is unprecedented in its collection and introduces a dataset for political bias and factuality for three tasks: *(i) media-level*, *(ii) article-level*, and *(iii) joint modeling at the article-level*. At the media and article levels, we evaluate the cross-lingual ability of the models; however, in joint modeling, we evaluate on English data. Our frameworks set a new benchmark in the cross-lingual evaluation of political bias and factuality. This is achieved via the use of various Multilingual Pre-trained Language Models (MPLMs) and Large Language Models (LLMs) coupled with ensemble learning methods.

1 Introduction

The integrity and objectivity of the news media are crucial in an age where information is rapidly disseminated across diverse cultural and linguistic landscapes (Fenton, 2009). As observed Vosoughi et al. (2018), misleading information or "fake news," spreads six times faster than the truth and reaches a much larger audience. This underscores the need for comprehensive data to assess political bias and factuality in news media and articles, particularly in a cross-lingual context, which remains a significant challenge (Nakov et al., 2024). Thus, we introduce a novel corpus **SAFARI** specifically designed for the cross-lingual analysis of political bias and factuality in news media and articles. Our work in developing this corpus is motivated by the absence of cross-lingual resources for detecting political bias and factuality in media and articles analysis. To address this issue, we offer a dataset for ten languages: (i) at *media-level* political bias, we have slightly less than 2k, and for factuality marginally over 2.6k media, (ii) at *article-level* we collect around 190k for political bias and around 190k of articles for factuality, and (iii) for *joint modeling* at article-level we have moderately less than 100k of English articles.

Furthermore, the methodology behind our study incorporates the use of MPLMs and LLMs to assess dataset tasks. Our approach enables MPLMs and LLMs to provide an evaluation of political bias and factuality across languages at the source and article levels and in joint modeling. Moreover, MPLMs and LLMs (using zero-shot learning) are coupled with ensemble learning methods for evaluation.

Our contributions are as follows:

- We introduce a data construction pipeline that delivers a large-scale corpus for cross-lingual evaluation of political bias and factuality, addressing both the media and the article levels. Also, we present an English-only dataset for the joint modeling assessment at the article-level.
- We evaluate and compare distant supervision *vs.* expert-annotated data at the article-level only for political bias.
- We employ MPLMs for analysis at the media-level, article-level, and in joint modeling leveraging ensemble learning, using hard and soft votings.
- We implement LLMs using zero-shot learning with Mistral$_{7B}$ (Jiang et al., 2023) and LLaMA2$_{7B}$ (Touvron et al., 2023) utilizing an ensemble approach based on hard voting.

[*]Equal contribution.
[†]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12217–12231
November 12-16, 2024 ©2024 Association for Computational Linguistics

In Section 2, we provide a review of previous and recent studies that focus on political bias, factuality, and joint modeling analysis. In Section 3, the data collection process and the subsequent examination are elaborated. In Section 4, the research tasks are defined and the statistics of the dataset are introduced, together with the frameworks and techniques. Section 5 delineates our analysis and the results obtained using multiple MPLMs with ensemble learning. Section 6 explores our investigation of zero-shot learning for our tasks using LLMs and nuances of distant supervision *vs.* expert annotated data. Finally, Section 7 summarizes our findings and suggests potential future directions.

2 Related Work

Datasets Predicting political bias and factuality in news media requires large-scale databases, with previous efforts like those of (Färber et al., 2020; Cremisini et al., 2019; Zubiaga et al., 2016; Hamborg et al., 2019; Kiesel et al., 2019a; Lim et al., 2020, 2018; Vargas et al., 2023) relied on crowdsourcing to gather data. However, these databases are relatively small and focus mainly on English, with annotations at the level of the media, article, and sentence (Baly et al., 2020a, 2018; Cremisini et al., 2019; Hamborg et al., 2019; Kiesel et al., 2019a). In contrast, our corpus, which emphasizes diverse languages, offers a larger dataset to predict political bias and factuality at the media and article levels. In the following, we explore the methods and datasets coupled to use political bias and factuality and their joint analysis.

Political Bias Understanding political bias is a nuanced exploration with varying definitions, including uneven coverage or favoritism (Stevenson et al., 1973) and systematic preferences for candidates or ideas (Waldman and Devitt, 1998). Guo et al. (2022) employed pre-trained BERT (Devlin et al., 2019) models to detect linguistic political bias in news articles. Groeling (2013) expanded the concept of media bias, considering dimensions such as selection and presentation of political bias influenced by the choice of newsmakers (Smith et al., 2001; Hassell et al., 2020). The study by Fan et al. (2019) used annotated media from Budak et al. (2016), analyzing 300 NYT, FOX, and HPO articles for bias, similar to our distant supervision approach to capture diverse ideological perspectives. Research on selection political bias requires huge databases, with studies using commercial (Soroka, 2012; Padgett et al., 2019; Gilens and Hertzman, 2000; Boykoff and Boykoff, 2004) and public datasets (Boudemagh and Moise, 2017; Kwak and An, 2014) using multi-source approaches (Kwak and An, 2016; Weaver and Bimber, 2008). Various methods measure the political bias of news media, including linking news outlets with politicians, analyzing shared audiences (Groseclose and Milyo, 2005; Gentzkow and Shapiro, 2010), and identifying the intricate linguistic techniques used to shape readers' opinions and emotions (Sajwani et al., 2024). Alternately, political bias is assessed through Twitter interactions (An et al., 2011, 2012; Stefanov et al., 2020). Predictions extend to political bias at the media, article and sentence levels, often using distant supervision with small datasets only in English (Kulkarni et al., 2018; Potthast et al., 2018; Kiesel et al., 2019b; Baly et al., 2020a; Da San Martino et al., 2023; Barrón-Cedeño et al., 2023a,b; Azizov et al., 2023; Chen et al., 2018; Fan et al., 2019; Spinde et al., 2022).

Factuality Veracity of information is examined at various levels: claim-level (e.g., "fact-checking"), article-level (e.g., "fake news" detection), user-level (e.g., identifying trolls), and medium-level (e.g., source reliability estimation). Claim-level efforts focus on fact-checking and rumor detection using social media interactions (Castillo et al., 2011; Canini et al., 2011; Ma et al., 2015; Ma et al., 2016, 2017; Kochkina et al., 2018; Dungs et al., 2018; Lim et al., 2020; Nguyen et al., 2020; Hardalov et al., 2022; Nakov et al., 2023), focusing on the stance and reliability of the source.

Early work estimated source reliability based on a medium's stance towards true/false claims using an English dataset (Mukherjee and Weikum, 2015; Dong et al., 2015; Popat et al., 2016, 2017; Popat et al., 2018). Recent approaches, such as Baly et al. (2020c), used gold labels and various English information sources, which are relatively small compared to our work. Mehta et al. (2022) and Panayotov et al. (2022) used graph-based frameworks to profile news media outlets, focusing on relationships and audience overlap. LLMs (e.g., ChatGPT) are used for the estimation of source reliability, as demonstrated by Yang and Menczer (2023), correlated with human expert ratings, and Mehta and Goldwasser (2023) introduced a framework that combined graph-based models, LLMs and human expertise for the profile of news media, effectively identifying fake news with minimal

Language	Political Bias						Factuality						
	Left	Left-Center	Least Biased	Right-Center	Right	Total	Very High	High	Mostly factual	Mixed	Low	Very Low	Total
English	259	567	637	279	134	1,876	67	1,529	166	425	202	119	2,508
German	-	9	5	6	1	21	1	8	8	3	1	2	23
Hindi	3	8	-	4	-	15	-	3	5	6	1	-	15
French	2	4	2	2	-	10	2	5	2	2	1	2	14
Spanish	1	3	2	3	-	9	-	7	2	3	-	-	12
Hebrew	1	2	1	2	2	8	-	5	-	6	-	-	11
Japanese	-	2	3	2	-	7	-	7	-	1	1	-	9
Italian	-	2	2	1	1	6	-	5	1	1	2	-	8
Arabic	-	3	1	1	1	6	-	3	-	3	1	-	7
Russian	-	-	-	2	-	2	-	-	-	2	2	2	6
Total						**1,960**							**2,613**

Table 1: Media-level dataset statistics.

	Political Bias				Factuality					
Language	Left	Center	Right	Total	Very High	High	Mixed	Low	Very Low	Total
English	51,076	52,939	34,801	138,816	8,661	56,656	13,838	12,937	4,095	96,187
Spanish	3,168	4,281	1,720	9,169	-	2,000	6,168	-	-	8,168
French	1,680	4,102	2,243	8,025	-	16,191	17,091	2,243	-	35,525
German	1,200	2,840	1,020	5,060	130	8,140	-	100	-	8,370
Italian	-	-	5,672	5,672	-	-	5,672	-	-	5,672
Bulgarian	-	4,860	-	4,860	-	-	4,860	-	-	4,860
Hindi	2,890	-	-	2,890	2,890	-	-	-	-	2,890
Persian	-	-	2,833	2,833	-	-	2,833	-	-	2,833
Polish	-	-	5,000	5,000	10,000	-	6,168	-	-	16,168
Russian	-	-	3,980	3,980	-	-	3,980	-	862	4,842
Total				**186,305**						**189,347**

Table 2: Article-level dataset statistics.

human input. Burdisso et al. (2024) employed reinforcement learning to estimate the reliability of the media, correlated with journalist scores to predict reliability labels.

Joint Modeling Joint modeling of factuality and political bias remains underexplored, with an attempt by (Baly et al., 2019) using a small English dataset using multi-task ordinal regression. Understanding the relationship between factuality and political bias in the news media, especially when outlets exhibit different behaviors on these aspects, presents a significant challenge.

3 Dataset Construction

Our methodology encompasses two levels of data collection: *media-level* and *article-level,* both using the distant supervision technique (Mintz et al., 2009) for article collection. We use a two-step criterion: (i) We exclusively used sources expertly annotated by **Media Bias/Fact Check**[1]. (ii) We select active media outlets. In *media-level,* we gather sources from Media Bias/Fact Check (MBFC) and collect up to 30 front-page articles from each website, labeled according to their sources. Similarly, in *article-level,* we apply distant supervision by assigning labels to articles from media annotated by MBFC, and collect expert-annotated data for political bias from **AllSides**[2] to compare performance with distant supervision data. In addition, during the data scraping process, we specifically targeted sections that focused on political, economic, and social issues. With this in mind, we used the EBK-means (Bholowalia and Kumar, 2014) clustering to analyze our entire dataset and identified 15 clusters. Furthermore, we validate our choice with the silhouette score, confirming the quality and separation of the clusters. The percentage distribution of data points was calculated across the clusters and visualized in Figure 1.

3.1 Media-level

Media Collection Figure 2 (Appendix A) shows our pipeline for media-level data collection, and the following are our steps: (i) At this stage, we compile a set of media sources from MBFC. After manually evaluating the availability of each source through their links, we extracted the details of each

[1] www.mediabiasfactcheck.com

[2] www.allsides.com

	Very High	High	Mixed	Low	Very Low	Total
Center	8,661	29,869	-	-	-	38,530
Left	-	26,787	2,587	-	-	29,374
Right	-	-	11,251	12,937	4,095	28,283
Total						**96,187**

Table 3: Joint modeling dataset statistics.

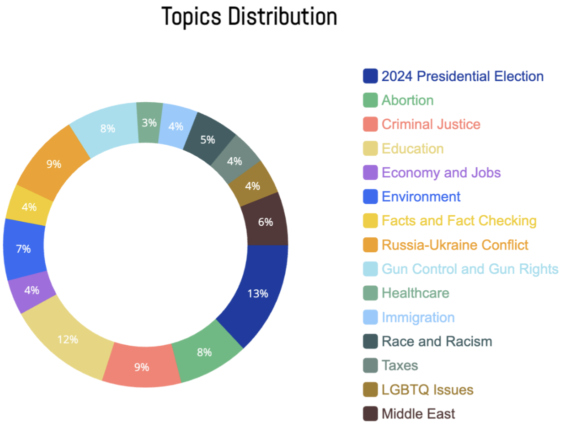

Figure 1: Topics distribution in our dataset.

source as JSON-formatted lines from the HTML code. (ii) In the article link parsing stage, front-page article links from these media sources were parsed according to specific criteria. Only links that were internal to the domain and have more than 65 characters in length, excluding links from the menu button. (iii) In the article text collection stage, the previously selected article links were used to retrieve the title and full text of the articles. We use script code and manually test to ensure effective text extraction. (iv) Finally, the post-processing stage involved formatting the collected data in the required JSON format.

3.2 Article-level

Articles Collection As illustrated in Figure 2 (Appendix A) for the data at the article-level we obtained the medium with the respective label from the data at the media-level. Subsequently, the selection of the media for parsing involved manually selecting the available sources with minimum 100 articles in their archive to have sufficient data to base our predictions on. Afterthat, it required to distinct structure of each website and analysis of their HTML code using a browser code inspector to identify relevant tags for efficient parsing. The articles parsing function facilitates this process in four stages: (i) initially retrieving the complete code from the archive page of the article, (ii) analyzing this code to extract a list of articles (including titles and links), (iii) making a secondary request to gather the full text of each article, and (iv) finally compiling these data into a JSON format.

Set	Media-level (A)	Media-level (B)	Article-level (A)	Article-level (B)	Joint modeling
Train	1,704	2,354	83,180	57,433	57,433
Development	86	77	10,000	10,000	10,000
Test (Eng)	86	77	28,180	28,754	28,754
Test (Multi)	84	105	47,489	54,527	
Test (Eng-EA)			17,456		

Table 4: Train/development/test sets distribution over media-level, article-level and joint modeling datasets.

Allsides Data obtained from AllSides were collected from the entire archive using a strategy similar to that used for the article-level.

3.2.1 Joint Modeling

To gather data on political bias and factuality at the article-level, for joint modeling, we utilized the method from the article-level as shown in Figure 2 (Appendix A), however, we combined the labels: political bias and factuality.

3.3 Data Curation

We applied the same curation method for media-level, article-level, and joint modeling. As shown in Figure 2 (Appendix A), the curation process involved evaluating the dataset according to the length of the article. Longer articles were typically found in sources considered more factual, while no similar trend was observed for political bias. To reduce the impact of very short or excessively long texts, which might be less relevant or contain mixed content (e.g., advertising), we focused on articles between 500 and 1,500 words. This range was chosen because the average article length in our dataset is 1,000 words. Although this approach may not entirely eliminate bias, it helps to ensure more informative representation and reduces potential bias across languages.

When we obtain media articles, we first remove duplicate content. We also meticulously removed HTML artifacts, such as tags, scripts, and CSS elements, to ensure that only actual textual content was retained. Alongside the advertisements, non-relevant elements such as navigation menus and footers were manually filtered out.

4 SAFARI Benchmark

4.1 Poltical Bias and Factuality

4.1.1 Media-level

(A) Given the news article(s) of a news outlet (e.g., www.bloomberg.com), predict the overall po-

Model	Hard Voting					Soft Voting				
	MAE	F1	A	P	R	MAE	F1	A	P	R
					English					
mBERT$_{\text{Base}}$	0.183	82.43	82.37	83.99	82.37	0.050	80.87	80.77	81.22	80.77
XLM-R$_{\text{Base}}$	0.215	79.80	79.71	80.79	79.71	0.128	81.59	81.46	81.22	81.46
mDeBERTaV3$_{\text{Base}}$	0.149	83.77	83.75	83.95	83.75	0.145	81.98	81.94	80.10	81.94
DistilmBERT$_{\text{Base}}$	0.176	83.64	83.78	87.23	87.78	0.125	84.46	84.37	84.19	84.37
mBART$_{\text{Large}}$	0.126	84.83	84.88	84.07	84.88	0.125	84.93	84.89	**85.08**	84.89
Ensemble	0.125	84.95	84.91	85.02	84.91	**0.120**	**84.96**	**84.92**	84.95	**84.92**
					Multilingual					
mBERT$_{\text{Base}}$	1.052	26.64	37.50	25.52	37.50	1.052	25.74	37.50	24.25	37.50
XLM-R$_{\text{Base}}$	1.062	26.54	36.45	25.77	36.45	1.104	23.58	32.29	22.03	32.29
mDeBERTaV3$_{\text{Base}}$	1.052	29.05	36.45	33.44	36.45	**1.010**	**32.12**	**39.58**	**40.26**	**39.58**
DistilmBERT$_{\text{Base}}$	1.302	22.98	26.04	21.82	26.04	1.364	20.29	22.91	19.46	22.91
mBART$_{\text{large}}$	1.063	27.45	33.33	37.72	33.33	1.062	27.02	33.32	37.17	33.32
Ensemble	1.117	27.44	37.62	27.14	37.62	1.118	26.88	36.63	25.52	36.63

Table 5: Analysis of political bias using hard and soft votings for each framework and ensemble at media-level (A). **Bold** values indicate the best scores for each category.

Model	Hard Voting					Soft Voting				
	MAE	F1	A	P	R	MAE	F1	A	P	R
					English					
mBERT$_{\text{Base}}$	**0.132**	**83.20**	**83.19**	**83.23**	**83.19**	0.090	82.93	82.59	82.50	82.59
XLM-R$_{\text{Base}}$	0.223	80.79	80.94	80.22	80.94	0.532	62.84	70.12	70.75	70.12
mDeBERTaV3$_{\text{Base}}$	0.188	81.82	81.99	81.03	81.82	0.207	81.22	81.01	81.63	81.01
DistilmBERT$_{\text{Base}}$	0.110	81.56	81.60	81.14	81.60	0.519	60.38	67.85	56.15	67.85
mBART$_{\text{Large}}$	0.049	82.39	82.38	82.41	82.38	0.415	71.28	64.15	82.15	64.15
Ensemble	0.142	81.49	81.59	81.15	81.59	0.143	81.83	81.36	81.48	86.36
					Multilingual					
mBERT$_{\text{Base}}$	1.183	29.60	27.50	36.60	27.50	0.980	30.25	35.57	31.01	35.57
XLM-R$_{\text{Base}}$	1.006	29.76	39.71	30.88	39.71	1.490	15.00	25.00	19.24	25.00
mDeBERTaV3$_{\text{Base}}$	1.054	24.78	30.37	38.52	30.37	1.230	21.34	27.88	37.35	27.88
DistilmBERT$_{\text{Base}}$	1.090	28.84	39.85	32.47	39.85	1.394	12.65	23.07	13.24	23.07
mBART$_{\text{Large}}$	1.386	25.45	22.73	35.70	22.73	1.240	27.00	29.80	29.91	29.80
Ensemble	0.872	38.44	50.00	44.18	50.00	**0.854**	**40.76**	**50.01**	**42.38**	**50.01**

Table 6: Analysis of factuality using hard and soft votings for each framework and in ensemble at media-level (B).

litical bias of that news outlet as: LEFT-, LEFT-CENTER, CENTER-, RIGHT-CENTER OR RIGHT-LEANING.

(B) Given the news article(s) of a news outlet (e.g., www.bloomberg.com), predict the overall factual reporting of that news outlet as: VERY HIGH, HIGH, MOSTLY FACTUAL, MIXED, LOW OR VERY LOW.

4.1.2 Article-level

(A) Given an article, classify its political bias as: LEFT, CENTER, OR RIGHT.

(B) Given an article, classify its factual reporting as: VERY HIGH, HIGH, MIXED, LOW, OR VERY LOW.

4.1.3 Joint Modeling

Given an article, classify its political bias and factual reporting jointly as: CENTER-VERY HIGH, CENTER-HIGH, LEFT-HIGH, LEFT-MIXED, RIGHT-MIXED, RIGHT-LOW AND RIGHT-VERY LOW.

Important In joint modeling of political bias and factuality, specific bias labels are strongly correlated with certain factuality levels (Baly et al., 2019). For example, a "center" bias typically corresponds to "very high" or "high" factuality. The expert-annotated data we collected from MBFC reflect this correlation, as it does not include uncommon combinations (e.g., left-low or right-high) shown in Table 3. This absence aligns with the source's correlation and annotation guidelines.

4.2 Dataset Statistics

4.2.1 Media-level

Table 1 presents the total amount of media and its distribution across languages for each label. For both sets, we have the same train/val/test sets. When data were acquired, as shown in Table 4, the dataset was segmented into training, development, and testing sets. There is a single combined training and validation set, exclusively in English. For testing, there are two distinct sets: the first is in English, while the second is multilingual for both political bias and factuality.

4.2.2 Article-level

A Table 2 presents the total number of articles with their political bias and distribution between languages for each label. Furthermore, Table 4 illustrates that we have a single set of training and validation articles, both exclusively in English, compiled using distant supervision. In addition, there are three testing sets: the first comprises English articles collected through distant supervision (DS), the second is an English test set assembled from AllSides, annotated by experts (EA), and the third is a multilingual test set of articles.

B As shown in Table 4, for the factuality of reporting of news articles, we have only one train and validation sets of articles in English. Our test sets comprise two distinct types: English and multilingual.

4.2.3 Joint Modeling

Table 3 presents the total number of articles and their distribution by label. Furthermore, Table 4 illustrates the distribution of data in train/validation and test sets, which are only given in English.

Note: We carefully split the dataset into train/development/test sets to avoid data leakage. Each split is unique and ensures that no media or articles previously exposed to the model are included in the other sets. The splits were performed using a stratified sampling approach to maintain the distribution of classes across all sets. The test sets are unique and exclude articles from sources previously exposed to the model. Moreover, the

	Political Bias					Factuality				
Model	MAE	F1	A	P	R	MAE	F1	A	P	R
					English-DS					
mBERT$_{\text{Base}}$	0.168	81.46	81.49	81.46	81.50	0.188	81.18	81.22	81.23	81.22
XLM-R$_{\text{Base}}$	0.130	81.33	81.37	81.35	81.37	0.160	81.54	81.57	81.55	81.57
mDeBERTaV3$_{\text{Base}}$	0.131	81.38	81.41	81.39	81.41	0.163	81.45	81.41	81.63	81.41
DistilmBERT$_{\text{Base}}$	0.162	81.23	81.27	81.25	81.27	0.162	81.49	81.52	81.49	81.52
Hard Voting	0.122	82.06	82.02	82.20	82.02	0.158	81.73	81.70	81.82	81.70
Soft Voting	**0.112**	**82.62**	**82.59**	**82.70**	**82.59**	**0.157**	**81.88**	**81.87**	**81.91**	**81.87**
					Multilingual					
mBERT$_{\text{Base}}$	0.630	61.60	61.26	67.41	61.26	0.492	66.54	66.75	68.31	66.75
XLM-R$_{\text{Base}}$	0.601	62.99	62.53	68.17	62.53	**0.479**	**67.22**	**67.67**	70.18	**67.67**
mDeBERTaV3$_{\text{Base}}$	0.609	62.85	62.42	66.68	62.42	0.480	66.07	67.59	**74.08**	67.59
DistilmBERT$_{\text{Base}}$	0.627	62.45	61.92	**69.32**	61.92	0.498	65.73	65.80	65.76	65.80
Hard Voting	**0.590**	**63.09**	**63.57**	66.97	**63.57**	0.497	65.89	65.81	65.99	65.81
Soft Voting	0.696	63.02	63.46	68.91	63.46	0.494	66.26	66.14	67.06	66.14
					English-EA					
mBERT$_{\text{Base}}$	0.223	67.33	67.38	68.70	67.38					
XLM-R$_{\text{Base}}$	0.228	66.86	66.95	68.36	66.95					
mDeBERTaV3$_{\text{Base}}$	0.229	67.52	67.32	68.96	67.32					
DistilmBERT$_{\text{Base}}$	0.233	66.47	66.54	67.80	66.54					
Hard Voting	0.200	69.44	69.46	69.39	69.46					
Soft Voting	**0.192**	**70.01**	**69.97**	**71.73**	**69.97**					

Table 7: Analysis of political bias and factuality using frameworks independently and ensembles using hard voting and soft voting at article-level. DS - distant supervision. EA - Expert annotated data from AllSides.

English and multilingual test samples are unique and have no connection between them, as they originate from different news outlets and languages. This separation ensures an unbiased evaluation of the model performance across different languages and contexts.

4.3 Cross-lingual Assessment

The dataset predominantly consists of data in English with labels for both tasks; however, dataset lacks labeled articles and media in some other languages. To address this challenge, we employ the cross-lingual assessment.

At the media-level, we employ five MPLMs: mBERT$_{\text{Base}}$$Base$ (Devlin et al., 2019), XLM-R$_{\text{Base}}$ (Conneau et al., 2019), DistilmBERT$_{\text{Base}}$ (Sanh et al., 2019), mDeBERTaV3$_{\text{Base}}$(He et al., 2021), and mBART$_{\text{Large}}$(Liu et al., 2020). However, at the article-level and in joint modeling, we used the same MPLMs with the exception of mBART.

In a previous study Baly et al. (2020a) to detect political bias at the article level, adversarial media adaptation and specially adapted triplet loss were used. Furthermore, to predict political bias and factuality at media-level Baly et al. (2018) utilized a comprehensive set of features extracted from various sources: articles, Wikipedia page, Twitter account, URL structure and web traffic data from target media and in joint modeling. Baly et al. (2019) investigates the detection of trustworthiness and political ideology in news outlets using a multi-task ordinal regression framework, establishing a connection between political bias and low trustworthiness. This research shows that joint modeling significantly exceeds isolated methods. In our study, we use a traditional ensemble learning method (Freund and Schapire, 1997) in the analysis at the article and media levels, using hard and soft votings for performance optimization; the architecture is shown in Figure 3 (Appendix A).

Model	MAE	F1	A	P	R
mBERT$_{\text{Base}}$	0.146	81.50	81.17	81.39	80.69
XLM-R$_{\text{Base}}$	0.147	81.35	82.82	81.23	80.44
mDeBERTaV3$_{\text{Base}}$	0.145	82.03	81.46	81.46	80.85
DistilmBERT$_{\text{Base}}$	0.149	81.01	82.50	83.15	80.06
Hard Voting	0.146	83.57	83.13	82.07	80.70
Soft Voting	**0.145**	**83.81**	**83.50**	**83.29**	**80.97**

Table 8: Analysis of politcal bias and factuality jointly using each model independently and in ensemble using hard and soft votings.

At the *media-level*, we integrate the predictions from individual articles into their media sources using both hard and soft voting methods, along with combining the models in an ensemble approach.

At the *article-level*, we collate multiple model predictions and individual models for classification of articles.

The elaboration of the hard voting is in Equation 1, and the soft voting is in Equation 2.

Let P_i be the predicted label political bias or factuality of the i-th article. The aggregated political bias and factuality P_m can be calculated as follows:

$$P_m = \text{mode}(P_1, P_2, \ldots, P_n). \quad (1)$$

For soft voting, let $P_{i,j}$ be the predicted probability of the i-th article belonging to the j-th political bias class or factuality. The aggregated political bias and factuality P_m can be calculated as follows:

$$P_m = \arg\max_j \left(\frac{1}{n} \sum_{i=1}^{n} P_{i,j} \right). \quad (2)$$

For *joint modeling*, our study uses One Hot Encoding (OHE) (Bishop, 2006) to accommodate multi-class labels (e.g., left, center, right; very high, high, mixed, low, very low) within our loss function. Our dataset comprises various classes representing different political biases and the factuality of the reporting. To effectively train our model, these classes are transformed into binary format, resulting in a label array such as $[0, 1, 0]$ for political bias and $[0, 0, 1, 0, 0]$ for factuality. This representation ensures an optimal interpretation by the model. Using OHE, we facilitate the model's ability to handle and learn from the multi-faceted

Model	Political Bias					Factuality				
	MAE	F1	A	P	R	MAE	F1	A	P	R
	English									
Mistral	1.247	**30.21**	33.67	37.19	33.67	**1.242**	11.12	21.07	**19.04**	21.07
LLaMA2	**1.134**	22.12	**34.70**	**40.34**	**34.70**	1.601	**19.10**	**26.21**	18.75	**26.21**
Ensemble	1.160	27.97	32.00	27.18	32.00	1.581	15.14	20.93	17.79	20.93
	Multilingual									
Mistral	1.564	**18.72**	**22.88**	17.54	**22.88**	**1.003**	7.99	20.03	**28.77**	20.03
LLaMA2	1.560	4.14	13.68	**36.20**	13.68	1.676	20.62	26.06	18.97	**26.06**
Ensemble	**1.484**	16.71	22.22	22.93	22.22	1.076	**25.73**	**30.81**	25.00	25.73

Table 9: Analysis of political bias and factuality of reporting using hard voting for each framework and ensemble of models at media-level.

nature of our data. We use the tokenizer's function in our pipeline, whose primary function is to convert textual data into embeddings, a critical step in preparing the data for model training. However, the tokenizer does not directly participate in the transformation of the label space. The conversion of label formats is handled by a separate function in our data pre-processing pipeline.

This task can be formulated as follows: Given the one-hot encoded vectors for political bias $\mathbf{y}_P$ and factuality $\mathbf{y}_F$, and the features $\mathbf{x}$ of an article, the joint prediction can be modeled as shown in Equation 3:

$$\hat{\mathbf{y}} = \text{softmax}(W\mathbf{x} + \mathbf{b}), \tag{3}$$

where $\hat{\mathbf{y}}$ is the predicted probability distribution over the joint classes of political bias and factuality, W is the weight matrix and $\mathbf{b}$ is the vector.

The loss function for training the model is defined as the sum of the cross-entropy losses for political bias and factuality, as expressed in Equation 4:

$$\mathcal{L} = -\left(\sum_j y_{P,j} \log \hat{y}_{P,j} + \sum_k y_{F,k} \log \hat{y}_{F,k} \right), \tag{4}$$

where $y_{P,j}$ and $y_{F,k}$ are the true labels of political bias and factuality, respectively, and $\hat{y}_{P,j}$ and $\hat{y}_{F,k}$ are the predicted probabilities.

Evaluation Measures We evaluate our frameworks using the following measures: Mean Absolute Error (MAE), F1 Score (F1), Accuracy (A), Precision (P), and Recall (R). We report MAE given the ordinal nature of both the factual and political bias classes (Baly et al., 2018, 2020b). Furthermore, we provide *Weighted Average* for F1, Precision and Recall due to class imbalance. Additionally, we evaluated the stability of our MPLMs by averaging the results over 3-5 independent runs using various seeds by computing the standard deviation.

Model	Political Bias					Factuality				
	MAE	F1	A	P	R	MAE	F1	A	P	R
	English-DS									
Mistral	**0.732**	45.06	48.70	**56.02**	48.70	1.637	13.99	21.30	15.15	21.30
LLaMA2	0.748	46.56	**48.92**	55.50	**48.92**	**1.233**	16.85	24.56	15.30	24.56
Ensemble	0.747	**46.84**	48.33	49.98	48.33	1.287	**20.72**	**27.54**	**18.24**	**27.54**
	Multilingual									
Mistral	0.880	40.62	42.26	45.22	42.26	1.744	10.46	19.31	14.87	19.31
LLaMA2	0.835	38.98	42.16	42.66	42.16	**1.581**	**16.19**	**23.19**	**14.96**	**23.19**
Ensemble	**0.841**	**43.30**	**44.41**	**44.89**	**44.41**	1.630	13.03	20.94	11.54	20.94
	English-EA									
Mistral	0.838	**40.05**	41.53	**43.50**	41.53					
LLaMA2	**0.809**	36.64	41.57	41.31	41.57					
Ensemble	0.817	39.67	**41.63**	42.28	**41.63**					

Table 10: Analysis of political bias and factuality using frameworks independently and ensembles using hard voting at article-level.

5 Experimental Setup & Results

5.1 Experimental Setup

The experimental setup for all tasks involved consistent hyper-parameters across various MPLMs, with minor task-specific adjustments. More details can be seen in Appendix A.

5.2 Results

Media-level In our analysis that includes the detection of political bias and factuality in various models, we observe a notable performance in English and multilingual contexts. For the detection of political bias, as illustrated in Table 5, the ensemble of models shines in the English set with higher scores, while mDeBERTaV3 excels in the data of multilingual political bias using soft voting. In contrast, DistilmBERT performs poorly in multilingual bias detection. When we analyze the factuality, as shown in Table 6, mBERT emerges as the best performer in the English dataset using hard voting, but XLM-R and DistilmBERT lag behind. In the multilingual context, soft voting outperforms others.

Article-level Analyzing political bias and factuality in English distant supervision and expert annotated sets, and multilingual set, the performance of various models and ensemble methods employing hard and soft voting reveals promising results, as shown in Table 7. In the English-DS context for both political bias and factuality, soft voting emerges as the most effective classifier, outperforming all individual models with the highest scores in all evaluation measures. For the multilingual test set, hard voting shows a slight advantage over

Model	MAE	F1	A	P	R
Mistral	0.351	29.68	29.68	12.12	29.68
LLaMA2	0.340	**31.84**	31.84	**27.88**	31.84
Ensmeble	**0.317**	23.62	**36.48**	18.15	**36.48**

Table 11: Analysis of political bias and factuality jointly using each model independently and in ensemble using hard voting.

other methods in detecting political bias. In contrast, XLM-R leads in the factuality assessment. In the English-EA dataset, only political bias is evaluated, and soft voting ensemble of models is the most effective.

Joint Modeling In analyzing the joint performance of political bias and factuality in multiple models and ensemble methods, we observe the distinction. According to the results in Table 8, the ensemble of models using soft voting clearly outperforms all other individual classifiers. However, a hard voting ensemble of models, slightly behind soft voting, while still showing good performance, especially in precision, where it almost matches soft voting. Among the individual models, mDeBERTaV3 is the most efficient in this joint task.

Summary In summary, our study reveals that employing the soft voting ensemble method is effective across all tasks, albeit with nuances. This effectiveness comes in part from soft voting by averaging scores, leading to performance variability depending on the balance of weak and strong models. This was particularly evident in the multilingual test sets for article-level political bias and factuality, as well as in the multilingual test set for media-level bias and the English test set for media-level factuality. Furthermore, given the time and cost constraints associated with human annotations, the use of distant supervision data is a helpful approach[3] (more details can be seen in Subsection 6.2). We observed that specific MPLMs, such as mBERT and XLM-R, excelled in different tasks. The media-level dataset includes up to 30 articles per media outlet, ensuring comprehensive training, although this results in a predominance of English data (around 95%). This predominance aids in transferring the model's predictive capabilities to other languages, but leads to lower performance compared to the article-level dataset, which is larger and offers more data for training. Furthermore, the performance discrepancy between the English and multilingual configurations, as shown in Tables 5 and 6, can be attributed to several factors. Despite using a multilingual pre-trained model, fine-tuning on English data does not generalize well to other languages due to differences in vocabulary, syntax, grammar, and cultural contexts. Additionally, the model may overfit to English-specific patterns due to intensive English training and insufficient exposure to diverse linguistic datasets during fine-tuning.

6 Discussion

In this section, our analysis focuses on the latest LLMs, specifically Mistral$_{7B}$ and LLaMA2$_{7B}$, examining their capabilities in zero-shot learning coupled with ensemble using hard voting.

Furthermore, we explore why models tested on distant supervision data exhibit higher performance levels compared to those tested on expert-annotated data, specifically regarding the detection of political bias at the article-level in the English language.

6.1 Overall Observation

A notable challenge in our study is managing text length, which poses complexities for LLMs. To mitigate this, we use BART (Lewis et al., 2019) for the summarization of English texts and mT5 (Xue et al., 2021) for the processing of multilingual content with a minimum text length of 128 and a maximum of 412. Our objective was to eliminate parsing artifacts, reduce the input length required by LLMs, enhance data quality, and accelerate inference time. Subsequently, the pre-processed texts were converted into task-specific prompts as outlined in Section 4 and fed into LLMs.

Based on our observation of the results in Tables 9 and 10, LLMs in zero-shot learning settings recognize political bias more effectively compared to factuality. Furthermore, due to the less fine-grained labels for political bias at article-level compared to factuality, LLMs easier predict political bias when there are fewer classes. In general, the performance of Mistral, LLaMA2, and their ensemble varies based on the tasks. Table 11 focuses

[3] We conducted a manual analysis of a total of 500 articles from 124 media outlets and 1000 articles from 219 media outlets, randomly selected from AllSides. We cross-referenced these articles with Media Bias/Fact Check labels. Interestingly, 471 (94.2%) and 945 (94.5%) of the articles aligned perfectly with their respective outlet label, demonstrating the reliability of the DS data for our tasks and strengthening our assumption. Furthermore, these articles were chosen to ensure a diverse representation of the dataset, covering various media sources and biases.

on joint modeling, where LLaMA2 outperforms Mistral, and hard voting stands out for its overall accuracy.

6.2 Distant Supervision vs. Expert Annotation

Two primary factors explain the performance difference between the models evaluated in EA *vs.* DS. First, the models were trained and evolved only on English data obtained via DS that differ in quality and detail from EA. Second, expert-annotated data, which are considered gold labels, are more accurate and have more detailed annotations. This complexity is a significant barrier for the models because, in their training and development phases, they have not been exposed to such data, making it difficult for them to appropriately identify and adjust to the nuances present in the expert-annotated test set.

7 Conclusion & Future Work

In this article, we introduce **SAFARI**, a new large-scale corpus for cross-lingual evaluation at the media and article levels, specifically designed for the detection of political bias and factuality of reporting, along with our data construction pipeline. Furthermore, we present an exclusive English dataset for joint modeling at the article-level. We also compare the performance of distant supervision *vs.* human-annotated data for political bias at the article-level. Moreover, our corpus is evaluated using MPLMs, and we implement hard and soft ensemble learning voting for all tasks. Lastly, we experimented with LLMs using hard voting.

In future work, our aim is to gather a larger multilingual corpus and conduct a more fine-grained analysis of political bias and factuality. Acknowledging that the U.S.-centric *left/center/right* political spectrum is not universally applicable, we plan to model biases that are more relevant to different regions and cultures. We also intend to collaborate with experts, seek alternative data sources, and expand the date ranges of news outlets to reduce data imbalance and create a larger and more diverse dataset. Furthermore, we plan to perform a multimodal analysis of political bias and factuality in news media and articles. We will also deepen our error analysis, breaking it down by language to improve performance. Additionally, we will conduct experiments to study cross-lingual abilities in detail, focusing on discrepancies in factuality and political bias for articles on the same topic across different languages, and stratify results based on topic distribution. Finally, we plan to investigate political bias and factuality using fine-tuned LLMs, potentially leveraging techniques such as LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023).

Limitations

We created a corpus for diverse languages, increasing the accessibility of NLP research in cross-lingual studies. However, we were only able to cover ten languages at the article and media levels, each. For some languages, we had only one or two labels assigned for both tasks due to the unavailability of annotated sources and articles in other languages. Additionally, for joint modeling, we intended to conduct a cross-lingual evaluation; however, we faced limitations in identifying sufficient media sources in other languages for an effective evaluation, primarily due to the challenge of finding comprehensive sources that encompass the necessary labels. Moreover, we find it problematic to use these data for news sites in some other countries. Furthermore, due to limited computational resources, we were unable to fully fine-tune our LLMs (e.g., Mistral and LLaMA2).

Ethical Statement & Bias

The dataset was compiled with a firm commitment to comply with legal and ethical standards. This involved a careful review of the terms of use of all websites and ensuring that data collection processes respect these terms. The compilation focused exclusively on publicly available data, without bypassing access control measures such as paywalls or subscription models. The data collection methods used were transparent and deliberately designed to minimize any potential adverse impact on the source websites. Including limiting the frequency of access to avoid any strain on their resources. The news articles are not publicly available; only the URLs of the media and the recipe scraping with labels are provided to support research while preserving the confidentiality of the source.

Users should consider inherent biases in the media sources and annotations when interpreting the results. We include a diverse range of media outlets to minimize potential bias. This dataset can exhibit certain label biases due to restricted domain coverage. However, we diligently worked to mitigate any detrimental biases by manual data assessment.

References

Jisun An, Meeyoung Cha, Krishna Gummadi, Jon Crowcroft, and Daniele Quercia. 2012. Visualizing media bias through Twitter. In *AAAI ICWSM*, volume 6.

Jisun An, Meeyoung Cha, P. Krishna Gummadi, and Jon Crowcroft. 2011. Media landscape in Twitter: A world of new conventions and political diversity. In *AAAI ICWSM*.

Dilshod Azizov, S Liang, and P Nakov. 2023. Frank at checkthat! 2023: Detecting the political bias of news articles and news media. *Working Notes of CLEF*.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020a. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020b. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020c. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Firoj Alam, Tommaso Caselli, Giovanni Da San Martino, Tamer Elsayed, Andrea Galassi, Fatima Haouari, Federico Ruggeri, Julia Maria Struß, Rabindra Nath Nandi, et al. 2023a. The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *European Conference on Information Retrieval*, pages 506–517. Springer.

Alberto Barrón-Cedeño, Firoj Alam, Andrea Galassi, Giovanni Da San Martino, Preslav Nakov, Tamer Elsayed, Dilshod Azizov, Tommaso Caselli, Gullal S Cheema, Fatima Haouari, et al. 2023b. Overview of the clef–2023 checkthat! lab on checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 251–275. Springer.

Purnima Bholowalia and Arvind Kumar. 2014. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Christopher M Bishop. 2006. Pattern recognition and machine learning. *Springer google schola*, 2:645–678.

Emina Boudemagh and Izabela Moise. 2017. News media coverage of refugees in 2016: a GDELT case study. In *AAAI ICWSM*.

Maxwell T Boykoff and Jules M Boykoff. 2004. Balance as bias: Global warming and the us prestige press. *Global environmental change*, 14(2):125–136.

Ceren Budak, Sharad Goel, and Justin M Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271.

Sergio Burdisso, Dairazalia Sánchez-Cortés, Esaú Villatoro-Tello, and Petr Motlicek. 2024. Reliability estimation of news media sources: Birds of a feather flock together. *arXiv preprint arXiv:2404.09565*.

Canini et al. 2011. Finding credible information sources in social networks based on content and social structure. In *IEEE SocialCom/PASSAT*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 675–684. ACM.

Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International conference on natural language generation*, pages 79–88.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Andres Cremisini, Daniela Aguilar, and Mark A Finlayson. 2019. A challenging dataset for bias detection: the case of the crisis in the ukraine. In *Social,* *Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, pages 173–183. Springer.

Giovanni Da San Martino, Firoj Alam, Maram Hasanain, Rabindra Nath Nandi, Dilshod Azizov, and Preslav Nakov. 2023. Overview of the CLEF-2023 CheckThat! lab task 3 on political bias of news articles and news media. In *Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum*, CLEF '2023, Thessaloniki, Greece.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-based trust: Estimating the trustworthiness of web sources. *VLDB Endow.*, 8(9):938–949.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3007–3014, New York, NY, USA. Association for Computing Machinery.

Natalie Fenton. 2009. News in the digital age. In *The Routledge companion to news and journalism*, pages 557–567. Routledge.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from U.S. daily newspapers. *Econometrica*, 78(1):35–71.

Martin Gilens and Craig Hertzman. 2000. Corporate ownership and news bias: Newspaper coverage of the 1996 telecommunications act. *The Journal of Politics*, 62(2):369–386.

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16.

Tim Groseclose and Jeffrey Milyo. 2005. A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Xiaobo Guo, Weicheng Ma, and Soroush Vosoughi. 2022. Measuring media bias via masked language modeling. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1404–1408.

Felix Hamborg, Anastasia Zhukova, and Bela Gipp. 2019. Automated identification of media bias by word choice and labeling in news articles. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 196–205. IEEE.

Momchil Hardalov, Anton Chernyavskiy, Ivan Koychev, Dmitry Ilvovsky, and Preslav Nakov. 2022. CrowdChecked: Detecting previously fact-checked claims in social media. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, AACL_IJCNLP '22, online.

Hans JG Hassell, John B Holbein, and Matthew R Miles. 2020. There is no liberal media bias in which news stories political journalists choose to cover. *Science advances*, 6(14):eaay9344.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019a. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019b. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Haewoon Kwak and Jisun An. 2014. A first look at global news coverage of disasters by using the gdelt dataset. In *SocInfo*, pages 300–308.

Haewoon Kwak and Jisun An. 2016. Two tales of the world: Comparison of widely used world news datasets GDELT and EventRegistry. In *AAAI ICWSM*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484.

Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. 2018. Understanding characteristics of biased sentences in news articles. In *CIKM workshops*, pages 121–128.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3818–3824. IJCAI/AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Ma et al. 2015. Detect rumors using time series of social context information on microblogging websites. In *CIKM*, pages 1751–1754.

Nikhil Mehta and Dan Goldwasser. 2023. An interactive framework for profiling news media sources. *arXiv preprint arXiv:2309.07384*.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23, 2015*, pages 353–362. ACM.

Preslav Nakov, Firoj Alam, Giovanni Da San Martino, Maram Hasanain, Rabindra Nath Nandi, Dilshod Azizov, and Panayot Panayotov. 2023. Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media. In *Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum*, CLEF '2023, Thessaloniki, Greece.

Preslav Nakov, Jisun An, Haewoon Kwak, Muhammad Arslan Manzoor, Zain Muhammad Mujahid, and Husrev Sencar. 2024. A survey on predicting the factuality and the bias of news media. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15947–15962, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. FANG: leveraging social context for fake news detection using graph representation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1165–1174. ACM.

Jeremy Padgett, Johanna L Dunaway, and Joshua P Darr. 2019. As seen on tv? how gatekeeping makes the US house seem more extreme. *Journal of Communication*, 69(6):696–719.

Panayot Panayotov, Utsav Shukla, Husrev Taha Sencar, Mohamed Nabeel, and Preslav Nakov. 2022. GREENER: Graph neural networks for news media profiling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP '22, Abu Dhabi, UAE.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *CIKM*.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the Web and social media. In *WWW Companion*, pages 1003–1012.

Popat et al. 2018. CredEye: A credibility lens for analyzing and explaining misinformation. In *The Web*.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Ahmed Sajwani, Alaa El Setohy, Ali Mekky, Diana Turmakhan, Lara Hassan, Mohamed El Zeftawy, Omar El Herraoui, Osama Mohammed Afzal, Qisheng Liao, Tarek Mahmoud, Zain Muhammad Mujahid, Muhammad Umar Salman, Muhammad Arslan Manzoor, Massa Baali, Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2024. FRAPPE: FRAming, Persuasion, and Propaganda Explorer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 207–213, St. Julians, Malta. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Jackie Smith, John D McCarthy, Clark McPhail, and Boguslaw Augustyn. 2001. From protest to agenda building: Description bias in media coverage of protest events in Washington, DC. *Social Forces*, 79(4):1397–1423.

Stuart N Soroka. 2012. The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, 74(2):514–528.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2022. Neural media bias detection using distant supervision with BABE–bias annotations by experts. *arXiv preprint arXiv:2209.14557*.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Stevenson et al. 1973. Untwisting the news twisters: A replication of Efron's study. *Journalism Quarterly*, 50(2):211–219.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. 2023. Predicting sentence-level factuality of news and bias of media outlets. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1197–1206, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Paul Waldman and James Devitt. 1998. Newspaper photographs and the 1996 presidential election: The question of bias. *Journal Mass Commun Q*, 75(2):302–311.

David A Weaver and Bruce Bimber. 2008. Finding news stories: a comparison of searches using Lexis-Nexis and Google News. *Journal Mass Commun Q*, 85(3):515–530.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kai-Cheng Yang and Filippo Menczer. 2023. Large language models can rate news outlet credibility. *arXiv preprint arXiv:2304.00228*.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11(3):1–29.

Appendix

A Data Statement for SAFARI

A.1 General Information

Dataset title SAFARI

Dataset version 1.0 (November 2023)

Data statement version 1.0 (October 2023)

Data collection period Media-level data were collected from July 2023 to September 2023. The article-level and joint modeling data were collected from September 2023 to November 2023. Articles span from September 2012 to November 2023.

A.2 Executive Summary SAFARI is a cross-lingual corpus focusing on ten languages at the media-level: English, German, Hindi, French, Spanish, Hebrew, Japanese, Italian, Arabic, and Russian. At the article-level, it includes English, French, Polish, German, Spanish, Italian, Bulgarian, Hindi, Persian, and Russian. Media Bias/Fact Check provided expert annotations for media-level data. Article-level data were collected from web archives of media outlets and supplemented with expert-annotated data from AllSides. Joint modeling used the article-level data collection approach.

Granularity The granularity of the analysis differs: the article-level uses a 3-point scale for political bias, while the media-level uses a 5-point scale. Media annotated as left-center and right-center were excluded to maintain distinct categories.

Difference in Media Counts Factuality annotations (2.6k) and political bias annotations (2k) differ due to the exclusion of sources labeled as "Questionable Source," "Conspiracy-Pseudoscience," or "Satire" in political bias, resulting in fewer total annotations.

A.3 Documentation for Source Datasets The SAFARI corpus was meticulously compiled for an in-depth analysis of political bias and factuality at the media and article levels and for joint modeling. At *media-level*, data was obtained from MBFC and annotated by experts. At *article-level*, data was collected directly from sources listed in the MBFC, with expert-annotated bias evaluations from AllSides. The *joint modeling* approach incorporated bias and factuality labels.

A.4 Language Variety The SAFARI corpus includes data in ten languages at both the media and article levels, but joint modeling includes only English.

Language Differences Data collection began at the media-level, followed by the article-level. Media outlets with fewer than 100 articles were excluded from the article-level dataset but retained in the media-level dataset, ensuring representation while maintaining a robust article-level dataset. Substitutions ensured at least 10 languages per task for cross-lingual analysis. Furthermore, MBFC annotations included the country of origin, which was manually verified before obtaining articles in the corresponding languages.

A.5 Experimental Setup

Hyper-parameters The *learning rate* was standardized to 2e-5 for all models: mBERT, XLM-R, DistilmBERT, mDeBERTaV3, and mBART. *Batch size* varied: 100 for mBERT and DistilmBERT, 80 for XLM-R, and 90 for mDeBERTaV3 and mBART. *Weight decay* and *maximum sequence length* were uniformly set at 0.01 and 512, respectively. During training, the model was validated every *100 steps* and saved every *15,000 steps*, with a limit of *three* checkpoints to manage storage.

Epoch Configuration Models were trained for 5 epochs at the media-level and 3 epochs at the article level and joint modeling data to prevent overfitting and enhance performance.

Hardware Our models were executed on NVIDIA RTX A6000 (48GB) GPU.

A.6 Library Selection We used `Requests` for retrieving page code and `BeautifulSoup (bs4)` for searching HTML elements. These libraries were chosen for their functionality and ease of use, with bs4's exception handling capabilities proving useful for parsing large datasets.

A.7 Cross-referencing AllSides with Media Bias/Fact Check Misaligned articles in cross-referenced subsets covered diverse topics. In a subset of 1000 articles, 7 out of 55 misaligned articles focused on "Trump," 4 on "Finance" and "Economy," and the rest on various topics like Elections, Criminal Justice, Environment, etc. In a subset of 500 articles, only 29 were misaligned, covering diverse topics.

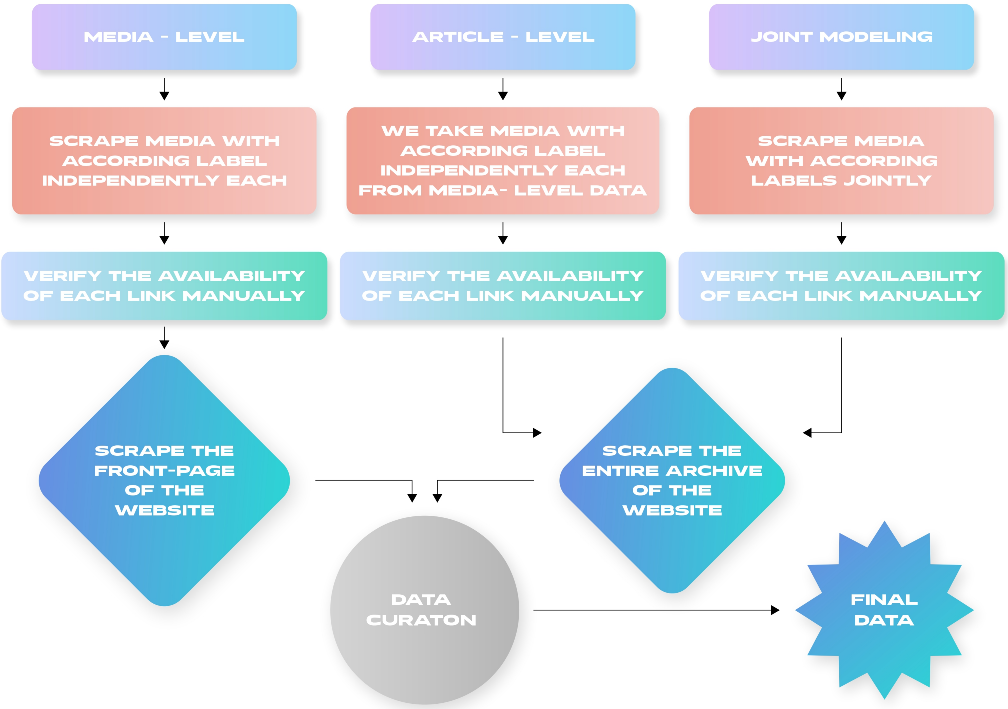
MEDIA - LEVEL
ARTICLE - LEVEL
JOINT MODELING
SCRAPE MEDIA WITH ACCORDING LABEL INDEPENDENTLY EACH
WE TAKE MEDIA WITH ACCORDING LABEL INDEPENDENTLY EACH FROM MEDIA- LEVEL DATA
SCRAPE MEDIA WITH ACCORDING LABELS JOINTLY
VERIFY THE AVAILABILITY OF EACH LINK MANUALLY
VERIFY THE AVAILABILITY OF EACH LINK MANUALLY
VERIFY THE AVAILABILITY OF EACH LINK MANUALLY
SCRAPE THE FRONT-PAGE OF THE WEBSITE
SCRAPE THE ENTIRE ARCHIVE OF THE WEBSITE
DATA CURATON
FINAL DATA

Figure 2: Data construction pipeline.

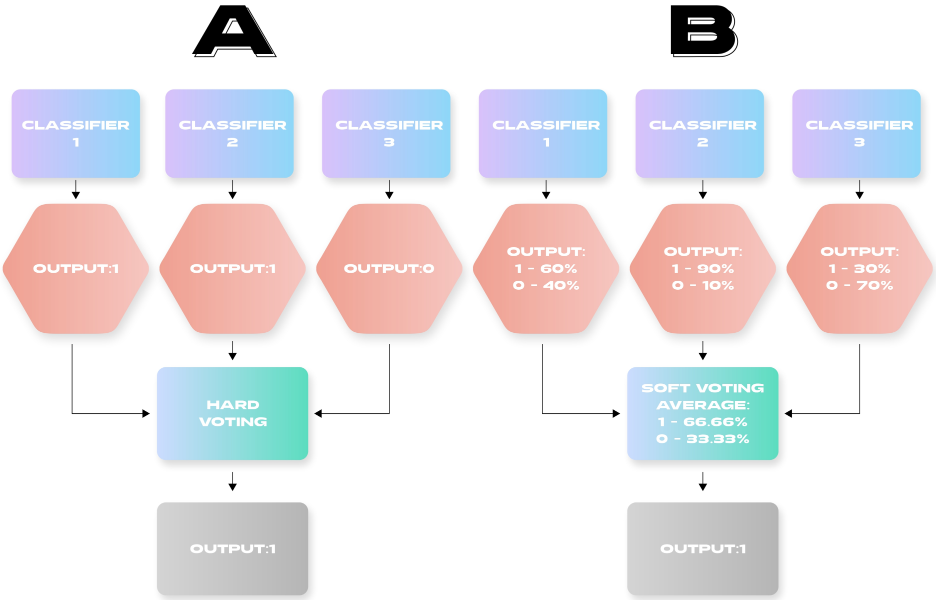

Figure 3: Architectures of hard voting (A) and soft voting (B) ensembles.