Profiling News Media for Factuality and Bias Using LLMs and the Fact-Checking Methodology of Human Experts

Zain Muhammad Mujahid[1,2] **Dilshod Azizov**[1] **Maha Tufail Agro**[1] **Preslav Nakov**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence, UAE
[2]University of Copenhagen, Denmark
{zain.mujahid, dilshod.azizov, maha.agro, preslav.nakov}@mbzuai.ac.ae

Abstract

In an age characterized by the proliferation of mis- and disinformation online, it is critical to empower readers to understand the content they are reading. Important efforts in this direction rely on manual or automatic fact-checking, which can be challenging for emerging claims with limited information. Such scenarios can be handled by assessing the reliability and the political bias of the source of the claim, *i.e.,* characterizing entire news outlets rather than individual claims or articles. This is an important but understudied research direction. While prior work has looked into linguistic and social contexts, we do not analyze individual articles or information in social media. Instead, we propose a novel methodology that emulates the criteria that professional fact-checkers use to assess the factuality and political bias of an entire outlet. Specifically, we design a variety of prompts based on these criteria and elicit responses from large language models (LLMs), which we aggregate to make predictions. In addition to demonstrating sizable improvements over strong baselines via extensive experiments with multiple LLMs, we provide an in-depth error analysis of the effect of media popularity and region on model performance. Further, we conduct an ablation study to highlight the key components of our dataset that contribute to these improvements. To facilitate future research, we released our dataset and code.[1]

1 Introduction

In an age where digital media dominate and information spreads quickly, profiling news media outlets in terms of their political bias and factuality is of utmost importance. Media organizations significantly shape public discourse (Sajwani et al., 2024), influence policy, and shape public opinion, making the detection of political bias essential (Pennycook and Rand, 2021).

Traditional methods for characterizing political bias in news media, such as subjective assessments and manual content analysis, are labor-intensive and prone to human biases. Automated content and social media analysis techniques (Baly et al., 2018, 2020b) have been developed for this, but they face limitations, including the laborious process of obtaining and annotating news articles.

Evaluating the factuality of the news reporting is equally important. Assessing the accuracy and truthfulness of news articles is the key to maintain the integrity of information dissemination (Baly et al., 2018). Conventional fact-checking methods are resource-intensive and struggle to keep up with the rapid production of news content. LLMs, such as the OpenAI GPT series (Radford et al., 2018, 2019; Brown et al., 2020), offer a promising solution. Trained on a vast amount of text datasets, LLMs can understand and generate human-like text, providing a new avenue for analyzing media bias and factuality at scale.

In this paper, we propose a novel methodology that leverages LLMs to predict political bias and the factuality of the reporting of entire news media outlets. Our approach (shown in Figure 1) involves crafting custom prompts to elicit responses from LLMs, thus enabling the detection of political bias and the factuality of a news outlet, without relying on the manual analysis of individual news articles. By employing the criteria used by professional fact-checkers, we aim to provide a more systematic and accurate assessment of political bias. We also conduct a case study to demonstrate the limitations of LLMs in the absence of well-defined guidelines, highlighting the importance of expert-driven prompts.

Furthermore, we conduct detailed error analysis to examine the impact of media popularity and region on the performance of the models, revealing biases in favor of more popular and U.S.-based outlets.

[1]https://github.com/mbzuai-nlp/llm-media-profiling

Findings of the Association for Computational Linguistics: ACL 2025, pages 798–819

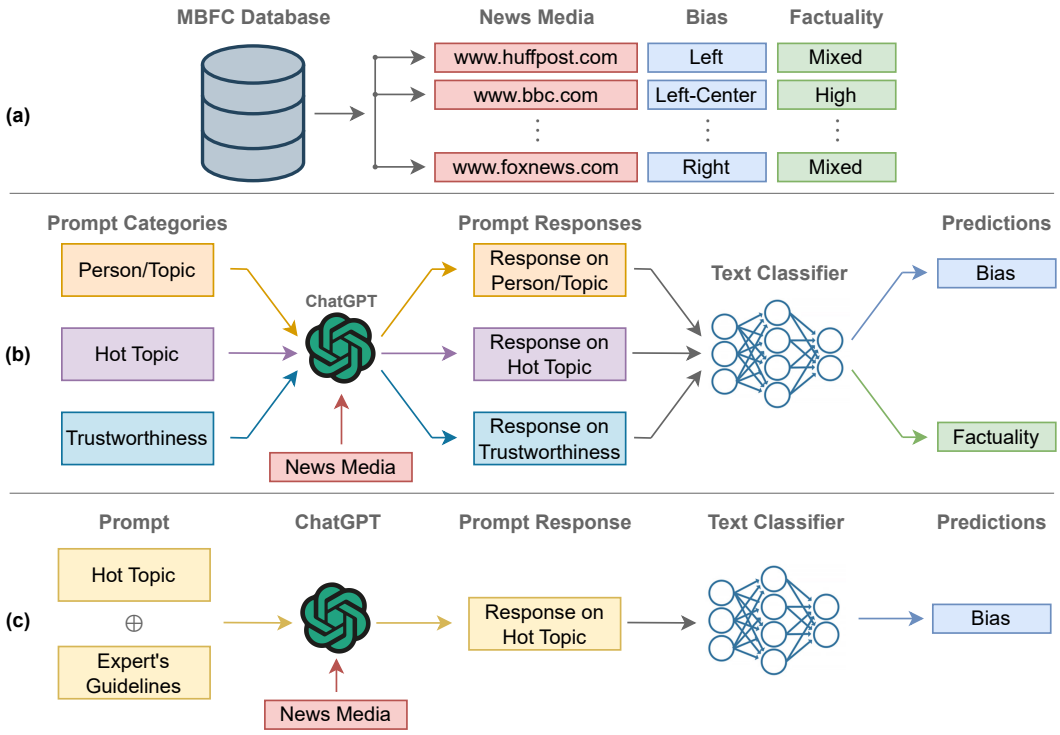

Figure 1: Overview of our methodology: (a) collection of gold labels from the MBFC[2] database, (b) data curation using handcrafted prompts, followed by text classification, (c) data curation using systematic prompts based on expert guidelines, followed by text classification.

We also perform an ablation study to identify key components of our dataset, demonstrating the importance of combining leaning and reasoning information for optimal results. This analysis not only highlights the strengths of our approach but also uncovers critical areas for future improvement. The following summarizes our key contributions:

- We release a large-scale dataset to model the factuality and political bias of news media.
- We leverage knowledge from LLMs to predict the factuality and the political bias of news media.
- We are the first to emulate the exact criteria used by professional fact-checkers when rating the political bias of the news media.
- We achieve sizeable improvements over baselines and zero-shot prompting for two tasks: predicting *(i)* the factuality of reporting and *(ii)* the political bias of news outlets.
- We conduct a comprehensive error analysis to examine the influence of media popularity and region on model performance.
- We perform an ablation study to evaluate the contribution of dataset components, showing that the reasoning extracted from LLMs is the most critical part for accurate predictions.

2 Related Work

The digital age has democratized the creation and dissemination of information via numerous media platforms, but this has also fueled misinformation (Naeem et al., 2021). News media profiling for political bias and factuality is essential to empower users and fact-checkers, ensure accountability, and support research (Nakov et al., 2024).

Political bias refers to systematic inclinations towards a candidate or ideology (Waldman and Devitt, 1998). Detecting political bias has been explored using a variety of features and methodologies, with predictive models operating across different levels of granularity, including media outlets, individual articles, and even sentences. For example, Baly et al. (2018) employed features from the NELA toolkit (Horne et al., 2018), while Kulkarni et al. (2018) examined article-level political bias by analyzing textual content and URLs, leveraging site-level annotations from AllSides[3]. At the outlet level, political bias can be detected by comparing media language to political speeches (Gentzkow and Shapiro, 2006).

[2] www.mediabiasfactcheck.com
[3] www.allsides.com

It can be also done by classifying articles along ideological axes such as left vs. right, or hyper-partisan vs. mainstream (Potthast et al., 2018; Saleh et al., 2019). Many of these models are based on distant supervision and are commonly trained on relatively small, English-language datasets (Da San Martino et al., 2023; Barrón-Cedeño et al., 2023; Barrón-Cedeño et al., 2023; Azizov et al., 2023, 2024).

Factuality prediction at the source level remains underexplored. Early research estimated the reliability of news sources by analyzing their stance on true or false claims rather than using explicit labels for medium-level factuality (Dong et al., 2016; Baly et al., 2019; Popat et al., 2016, 2018). Baly et al. (2018) explored political bias and factuality by extracting features from news articles, Wikipedia entries, Twitter metadata, and URLs, showing that integrating these sources improved classification accuracy. Later, Baly et al. (2019) found that the joint prediction of political bias and factuality was more effective. Baly et al. (2020a) used both the linguistic aspects and the social context. This included analyzing the text of articles, audio content, and social media reactions and discussions on platforms such as Facebook, Twitter, and YouTube, as well as Wikipedia content about the medium. Further, Azizov et al. (2024) examined a cross-lingual evaluation of political bias and factuality.

Also, some research has focused on developing LLMs such as GPT (Brown et al., 2020) and ChatGPT, demonstrating versatility in general-purpose reasoning tasks, including assessing factual accuracy and detecting political bias (Qin et al., 2023). Yang and Menczer (2025) evaluated ChatGPT's ability to gauge news outlet credibility across domains, including non-English and satirical sources, finding a moderate correlation with human expert evaluations (Spearman's $\rho = 0.54$, $p < 0.001$). Mehta and Goldwasser (2024) proposed an interactive framework combining graph-based models, LLMs, and human input to profile news sources and identify biased content. Although effective, Manzoor et al. (2025) proposed an approach to profile news media by integrating graph neural network representations with pre-trained language models, significantly boosting performance. Wang et al. (2024) addressed concerns about the factual accuracy of LLM outputs, proposing solutions for annotating LLM-generated responses.

Unlike the above studies, we do not use LLM-generated credibility ratings or human labor. Instead, we use questions on various levels of factuality and political bias, prompting LLMs to gather insights based on their internal parametric knowledge, which we then aggregate to make predictions.

3 Methodology

To predict the political bias and the factuality of a news outlet, we use a data-gathering technique where we prompt the LLM with targeted queries on various trendy or controversial topics. This method operates under the assumption that LLMs already encompass a vast amount of data (Yang et al., 2024) and possess knowledge that would allow them to profile media outlets effectively. We use the OpenAI `Python` package to query the `API` endpoint of the *gpt-3.5-turbo-0125* (GPT-3.5) (Ouyang et al., 2022; Brown et al., 2020), released on January 25, 2024. Specifically, we formulate two approaches, shown in Figure 1: first, we manually craft 18 questions to query the LLM regarding political bias and the factuality of a media outlet. Second, we integrate expert guidelines into our prompts to obtain improved responses to classify the political bias of the outlet. These LLM responses are concatenated and passed to text classification models. We also present two case studies where we obtain zero-shot predictions from the LLMs by providing the media name and some of its recently published articles.

3.1 Handcrafted Prompts

For this approach, we design a set of 18 questions to query the LLM. These questions are grouped into three categories (described below). Each question is asked in a separate `API` call to mitigate and minimize any potential bias of an ongoing session, which is standardized across our study.

3.1.1 Category 1: Stance on Public Figures or Topics

This category includes seven questions that aim to determine a website's stance on a public figure, personality, or topic. We use the following prompt:

```
system_prompt = "You are an assistant tasked with
↪  determining the stance of a website on a
↪  public figure personality or a topic."
user_prompt = "Is {domain}
↪  pro-{topicPublicFigure}. The assistant must
↪  provide the stance and reasoning in JSON
↪  format, and should not provide anything
↪  else."
```

The **{domain}** and **{topicPublicFigure}** placeholders in the prompts are dynamically replaced with the relevant domain (*e.g.,* foxnews.com) and topics or public figures (*e.g., Trump*, *Biden*, *Putin*, *American*, *Russian*, *German*, or *Palestinian*) to analyze media's stance and determine the political bias. A sample GPT-3.5 response for foxnews.com can be seen in the Appendix A.1.

3.1.2 Category 2: Stance on Current Popular Topics

In this category, we present five questions that aim to determine the stance of a news source on a trending topic. We use the following prompt:

```
system_prompt = "You are an assistant to help the
↪  user determine the stance of a website on a
↪  certain topic."
user_prompt = "What is the stance of {domain} on
↪  {topic}. The assistant should return the
↪  stance and the reason in JSON format, and
↪  should not provide anything else."
```

The **{domain}** placeholder represents the target domain, and **{topic}** include options like *Ukraine*, *Climate Change*, *Gun Control*, *Immigration*, or *Abortion*. A sample GPT-3.5 response for foxnews.com is given in the Appendix A.2.

3.1.3 Category 3: Factuality of the Reporting

This category includes six questions aimed at determining the factuality of a news outlet's reporting. We use the following prompt:

```
system_prompt = "You are an assistant to help the
↪  user determine the credibility of a website."
```

Each user prompt incorporates a common **{default_text}** directive as shown below:

```
default_text = "Assistant must return the
↪  'answer' and 'reason' in JSON format, and do
↪  not provide anything else."
user_prompt_q1 = "Can I trust {domain}. Answer in
↪  1 word. {default_text}"
user_prompt_q2= "Has {domain} failed fact-checks?
↪  Answer in one word. {default_text}"
user_prompt_q3 = "Is {domain} LEFT, CENTER, or
↪  Right. Answer in one word. {default_text}"
user_prompt_q4 = "Is {domain} Biased.
↪  {default_text}"
user_prompt_q5 = "Characterize {domain} in 3
↪  words. {default_text}"
user_prompt_q6 = "Characterize the bias of
↪  {domain} in 3 words. {default_text}"
```

The **{domain}** placeholder represents the target domain. A sample GPT-3.5 response for foxnews.com is given in the Appendix A.3.

3.2 Systematic Prompts

The second method adopts a systematic approach to querying an LLM for profiling news media political bias, leveraging the methodology employed by fact-checking journalists from Media Bias/Fact Check (MBFC)[4]. This methodology rates editorial political bias across 16 policy areas: *General Philosophy*, *Abortion*, *Economic Policy*, *Education Policy*, *Environmental Policy*, *Gay Rights*, *Gun Rights*, *Health Care*, *Immigration*, *Military*, *Personal Responsibility*, *Regulation*, *Social View*, *Taxes*, *Voter ID*, and *Worker's/Business Rights*. The left- and right-wing definitions for these topics are detailed in the Appendix B.1. In our LLM prompt below, the **{topic}** placeholder represents one of the 16 policy areas, while **{defLeft}** and **{defRight}** are replaced with their respective definitions. The **{topic}** placeholder is replaced by the domain being queried.

```
system_prompt = '''You are an expert journalist
↪  tasked with determining the bias leaning of a
↪  media outlet on the topic, {topic}. Return
↪  the 'leaning' and the 'reason' in JSON
↪  format. Consider your training data as
↪  evidence. Return 'unknown' if you cannot
↪  determine the leaning.
Labels Definition:
    - left: {defLeft}
    - right: {defRight}'''
user_prompt = "{domain}"
```

A sample GPT-3.5 response for foxnews.com is given in the Appendix B.2. Since expert definitions are only available for left and right labels, this method is not used for experiments in the 5-class political bias setting.

4 Case Study: Zero-Shot Predictions

We evaluate zero-shot capabilities of several LLMs, including Mistral 7B (Jiang et al., 2023), LLaMA2 7B, LLaMA3 (8B and 70B) (Touvron et al., 2023) and GPT-3.5 (Ouyang et al., 2022; Brown et al., 2020). Predictions for political bias and factuality are obtained using two methods: (a) providing only the media outlet name and (b) providing recent articles from the media along with the media name. We describe both approaches in § 4.1 and § 4.2.

[4] www.mediabiasfactcheck.com/methodology/

4.1 LLM Prompt Given a Media Name

The following is the prompt we used, where we only provided the media name as the input to obtain predictions about the factuality of the reporting:

```
system_prompt = '''You are an expert in media
↪  analysis. Classify the factual reporting
↪  level of a media given its name ONLY from one
↪  of three categories from the list provided
↪  below:
- high
- mixed
- low
Return -1 if you can not classify.'''
user_prompt = "{domain}"
```

The **{domain}** placeholder is replaced with the domain of interest. If the LLM is unable to determine the label or the outlet is unknown, it admits its uncertainty and returns -1, an example output of which is shown in Appendix C.1. The prompt used to predict political bias is given in Appendix C.2. For fine-grained predictions, we add two more classes, *left-center* and *right-center*, in this prompt.

4.2 LLM Prompt Given Media Articles

For this case study, we collected five articles from each media outlet, focusing on political, economic, and social issues. Due to the small number of articles, they were not divided across different timeframes. Each article was summarized using LLaMA3-70B with the prompt described in the Appendix C.3. These summaries served as evidence, which we attached to our prompt and asked LLMs to predict the factuality based on this evidence as follows:

```
system_prompt = '''You are an expert in media
↪  analysis. Classify the overall factual
↪  reporting level of the given news article of
↪  {media} into one of three categories from the
↪  list provided below:
- high
- mixed
- low
Return -1 if you can not classify.'''
user_prompt = "{article}"
```

The **{domain}** placeholder is replaced with the domain of interest, while the **{article}** placeholder is replaced with the summarized article. The prompt used to predict political bias is given in the Appendix C.4. For fine-grained predictions, we include two additional classes, *left-center* and *right-center*, in this prompt.

For each media outlet, we applied hard-voting from five predictions to assign a final label for evaluation. The summarized articles provided context for LLMs, improving accuracy, while the hard-voting technique ensured that the final label reflects consensus from multiple pieces of evidence. This reduces variability and potential bias from individual articles, offering a more balanced assessment of a media outlet's political bias and factuality.

5 Experiments and Evaluations

5.1 Dataset

To evaluate our system, we use the political bias and factuality labels provided by MBFC. An example annotation for `cnn.com` is shown in the Appendix D. Factuality is assessed on a three-point scale: *low*, *mixed*, and *high*. Political bias was originally modeled on a seven-point scale, but previous research (Baly et al., 2020a; Panayotov et al., 2022) simplified it to a three-point scale (*left*, *center*, and *right*), which we adopt for consistency with prior work. Table 5 in the Appendix D presents the label distribution in our dataset, which is larger and more granular than previous datasets, including fringe labels, such as *left-center* and *right-center*.

5.2 Experimental Setup

We use data collected in § 3.1 & 3.2 to train our models. Initially, the data is vectorized using TF-IDF to train an SVM classifier for the prediction of political bias and factuality. A grid search is conducted to tune C and γ for the RBF kernel.

For experimentation with transformer-based models, we use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019) separately for each task, i.e., political bias and factuality. Fine-tuning is performed with a $1e-5$ learning rate, batch size 16, dropout 0.2, over 5 epochs on NVIDIA RTX A6000 48GB. We maintain a train/test split of 80/20 for all of our experiments with a fixed seed value.

We compare our results with the majority class baseline and zero-shot prompting techniques using LLaMA, Mistral, and GPT-3.5. Two scenarios are tested: *(i)* predicting political bias and factuality using only the media name (§4.1), serving as an ablation study without information retrieval, and *(ii)* adding articles as evidence in the prompt (§4.2), to compare our methodology with traditional article-based profiling.

Class →	Low			Mixed			High			Acc.↑	MAE↓
Model ↓	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1		
Majority Class Baseline											
Majority class	.000	.000	.000	.000	.000	.000	.571	1.000	.727	.571	.572
Zero-Shot Baselines: LLM Prompt Given Name of Media											
GPT-3.5$_{\text{Turbo}}$	.260	.708	.380	.367	.362	.365	.931	.534	.679	.510	.619
Mistral-7B$_{\text{Instruct-v0.1}}$	.181	.233	.204	.288	.678	.404	.750	.187	.300	.335	.753
LLaMA2-7B$_{\text{Chat}}$	.200	.276	.232	.292	.661	.405	.763	.208	.327	.348	.744
LLaMA3-8B$_{\text{Chat}}$	.333	.275	.301	.274	.635	.383	.560	.213	.309	.343	.726
LLaMA3-70B$_{\text{Chat}}$	.473	.792	.592	.352	.471	.403	.839	.555	.668	**.565**	**.471**
Zero-Shot Baselines: LLM Prompt Given Articles from Media											
GPT-3.5$_{\text{Turbo}}$	.508	.882	.645	.812	.394	.531	.600	.455	.517	.580	.610
Mistral-7b$_{\text{Instruct-v0.1}}$	.324	.353	.338	.259	.212	.233	.167	.182	.174	.250	1.040
LLaMA2-7B$_{\text{Chat}}$	.333	.353	.343	.303	.303	.303	.258	.242	.250	.300	.940
LLaMA3-8B$_{\text{Chat}}$	.345	.294	.317	.291	.485	.364	.438	.212	.286	.330	.780
LLaMA3-70B$_{\text{Chat}}$	.705	.912	.795	.586	.515	.548	.741	.606	.667	**.680**	**.360**
Our Method (Hand-Crafted Prompts)											
SVM$_{\text{TF-IDF}}$	.736	.650	.690	.685	.671	.678	.878	.912	.895	**.806**	**.206**
BERT$_{\text{Base}}$	.629	.650	.639	.683	.575	.624	.858	.919	.887	.782	.238
RoBERTa$_{\text{Base}}$	.658	.642	.650	.676	.608	.640	.874	.923	.897	.793	.219
DistilBERT$_{\text{Base}}$	.672	.650	.661	.668	.629	.648	.875	.908	.891	.791	.222

Table 1: Results for *factuality* prediction. **Bold** values indicate the best scores for each category.

Class →	Left			Center			Right			Acc.↑	MAE↓
Model ↓	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1		
Majority Class Baseline											
Majority class	.000	.000	.000	.427	1.000	.598	.000	.000	.000	.427	.573
Zero-Shot Baselines: LLM Prompt Given Name of Media											
GPT-3.5$_{\text{Turbo}}$	.398	.537	.457	.744	.699	.721	.685	.614	.648	.636	.497
Mistral-7B$_{\text{Instruct-v0.1}}$	.882	.750	.811	.824	.945	.880	.927	.843	.883	.869	**.152**
LLaMA2-7B$_{\text{Chat}}$	.870	.750	.805	.823	.940	.878	.927	.843	.883	.867	.154
LLaMA3-8B$_{\text{Chat}}$	.318	.762	.449	.714	.192	.303	.727	.819	.771	.542	.540
LLaMA3-70B$_{\text{Chat}}$	.855	.738	.792	.877	.934	.905	.878	.873	.875	**.874**	.168
Zero-Shot Baselines: LLM Prompt Given Articles from Media											
GPT-3.5$_{\text{Turbo}}$	.596	.848	.700	1.000	.647	.786	.742	.697	.719	.730	.420
Mistral-7B$_{\text{Instruct-v0.1}}$	.310	.273	.290	.352	.576	.437	.235	.118	.157	.320	.870
LLaMA2-7B$_{\text{Chat}}$	.421	.485	.451	.391	.545	.456	.500	.235	.320	.420	.730
LLaMA3-8B$_{\text{Chat}}$	.290	.545	.379	.556	.147	.233	.448	.394	.419	.360	.950
LLaMA3-70B$_{\text{Chat}}$	.722	.788	.754	.800	.706	.750	.735	.758	.746	**.750**	**.340**
Our Method (Hand-Crafted+Systematic Prompts)											
SVM$_{\text{TF-IDF}}$	.914	.800	.853	.915	.940	.927	.883	.910	.896	.902	.133
SVM$_{\text{TF-IDF}}$†	1.000	.850	.919	.859	.962	.907	.942	.886	.913	.911	.093
BERT$_{\text{Base}}$	.859	.762	.808	.887	.902	.894	.884	.916	.899	.881	.147
BERT$_{\text{Base}}$†	.949	.925	.937	.908	.967	.937	.962	.904	.932	**.935**	**.075**
RoBERTa$_{\text{Base}}$	.827	.775	.800	.918	.913	.915	.901	.934	.917	.895	.138
RoBERTa$_{\text{Base}}$†	.923	.900	.911	.877	.934	.905	.936	.880	.907	.907	.103
DistilBERT$_{\text{Base}}$	.797	.787	.792	.885	.885	.885	.892	.898	.895	.872	.159
DistilBERT$_{\text{Base}}$†	.912	.912	.912	.862	.923	.892	.928	.855	.890	.895	.114

Table 2: Results for *political bias* prediction (3-point scale). Each model marked with the † symbol indicates that it is trained on data derived from prompts incorporating expert guidelines. **Bold** values indicate the best scores for each category.

Evaluation Measures: We use class-wise F1-score along with overall accuracy. We also report Mean Absolute Error (MAE) to account for the ordinal nature of the classes (Baly et al., 2020b,a; Azizov et al., 2024).

5.3 Factuality Prediction

Table 1 reports the evaluation results for the experiments on our dataset for predicting the factuality of the reporting of the news media, grouped by different modeling methodologies.

We observe that converting our gathered data from LLMs into embeddings using the TF-IDF vectorizer and training an SVM on it yields better results than any other approach in the table. This method achieves a final accuracy of 80.6% and the lowest MAE score of 0.206, indicating high precision and reliability in predicting the factuality.

In contrast, when we fine-tune transformer-based models using data gathered by prompting LLMs, we find that their accuracies are lower compared to the top-performing SVM model. The likely reason for this is that SVMs, combined with TF-IDF, effectively handle sparse, high-dimensional data and perform well with smaller datasets.

In our zero-shot experimentation, when given the media name and asked to predict its label, we observe that LLaMA2-7B, Mistral-7B, and LLaMA3-8B perform poorly. Their high MAE values indicate frequent misclassification between high and low labels. GPT-3.5 and LLaMA3-70B perform somewhat better, with LLaMA3-70B achieving the best accuracy of 56% in this category, along with the highest recall, suggesting its better knowledge.

However, the overall MAE in this category suggests that all models struggle to detect the exact labels accurately. This implies that providing the media name is insufficient to extract accurate information from the LLM, highlighting the need for more robust approaches to leverage LLMs effectively.

We observe a similar performance trend when we attach articles from the respective media to the prompt and ask the model to predict the factuality using these articles as evidence. As described in § 4.2, hard voting increases accuracy by approximately 12% compared to providing only the media name, with LLaMA3-70B achieving the best performance: 68% accuracy. Smaller models such as Mistral-7B, LLaMA2-7B, and LLaMA3-8B continue to struggle, having very high MAE scores, while GPT-3.5 performs better than these smaller models. This case study demonstrates that including articles in the prompt results in more confident and accurate responses from LLMs, as they can use the provided evidence to reason about their final label. However, our methodology, which uses handcrafted prompts, outperforms this technique, highlighting the importance of prompt design in detecting the factuality of the reporting of the news media.

Class →	Left			Left-Center			Center			Right-Center			Right			Acc.↑	MAE↓
Model ↓	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1	Pre.	Rec.	F1		
Majority Class Baseline																	
Majority class	.000	.000	.000	.000	.000	.000	.256	1.000	.407	.000	.000	.000	.000	.000	.000	.256	1.068
Zero-Shot Baselines: LLM Prompt Given Name of Media																	
GPT-3.5$_{\text{Turbo}}$	.564	.713	.630	**.576**	.320	.412	.443	.448	.446	.388	.326	.354	.567	.819	.670	.502	.672
Mistral-7B$_{\text{Instruct-v0.1}}$	.552	.662	.602	.676	.150	.246	.384	.776	.514	.475	.161	.241	.685	.830	.751	.505	**.639**
LLaMA2-7B$_{\text{Chat}}$	.552	.662	.602	.697	.150	.247	.382	.776	.512	.475	.161	.241	.688	.830	.753	.505	.640
LLaMA3-8B$_{\text{Chat}}$	.127	.674	.214	.308	.080	.127	.444	.094	.155	1.000	.087	.160	.390	.488	.433	.243	1.709
LLaMA3-70B$_{\text{Chat}}$	.411	.725	.525	.595	.307	.405	.415	.798	.546	.692	.050	.093	.729	.782	.754	**.510**	.727
Zero-Shot Baselines: LLM Prompt Given Articles from Media																	
GPT-3.5$_{\text{Turbo}}$	.349	.750	.476	.333	.200	.250	.526	.500	.513	.286	.100	.148	.789	.750	.769	**.460**	**.930**
Mistral-7B$_{\text{Instruct-v0.1}}$	.318	.350	.333	.222	.200	.211	.150	.150	.150	.389	.350	.368	.182	.200	.190	.250	1.500
LLaMA2-7B$_{\text{Chat}}$	.100	.100	.100	.227	.250	.238	.278	.250	.263	.200	.150	.171	.200	.250	.222	.200	1.710
LLaMA3-8B$_{\text{Chat}}$	.148	.450	.222	1.000	.150	.261	.800	.200	.320	.500	.200	.286	.348	.400	.372	.280	1.850
LLaMA3-70B$_{\text{Chat}}$	.362	.850	.507	.400	.200	.267	.471	.400	.432	.250	.100	.143	.778	.700	.737	.450	.960
Our Method (Hand-Crafted Prompts)																	
SVM$_{\text{TF-IDF}}$	.771	.675	.720	.482	.342	.400	.574	.721	.639	.628	.597	.612	.811	.849	.830	.648	.475
BERT$_{\text{Base}}$	.639	.754	.692	.483	.389	.431	.700	.744	.721	.717	.630	.670	.810	.914	.859	**.700**	.425
RoBERTa$_{\text{Base}}$	.704	.820	.758	.491	.481	.486	.683	.662	.672	.651	.651	.651	.863	.853	.858	.689	**.405**
DistilBERT$_{\text{Base}}$	.681	.803	.737	.467	.324	.383	.663	.708	.685	.616	.646	.630	.859	.859	.859	.676	.427

Table 3: Results for *political bias* prediction (5-point scale). **Bold** values indicate the best scores for each category.

5.4 Political Bias Prediction

Table 2 shows results for the political bias prediction task, grouped by modeling methods. Models marked with † were trained on data from systematic prompts described in § 3.2. This data improves accuracy: the SVM trained on vectorized data from systematic prompts outperforms the one using handcrafted prompts. Fine-tuned models also benefit, with BERT achieving the highest accuracy of 93.50%, outperforming all baselines. These results suggest that incorporating expert definitions into prompts helps elicit more accurate and confident predictions from LLMs.

Our experimental results for predicting political bias on a five-point scale can be seen in Table 3. We observe that transformer models perform the best when fine-tuned on the data gathered using handcrafted prompts, surpassing other models, including the majority class baseline and the SVM trained on TF-IDF vectorized text. BERT achieved the highest accuracy at 70%, while using RoBERTa yields the best MAE score of 0.405, indicating the least confusion among the ordinal classes. This shows that transformer-based models are highly efficient in modeling political bias beyond three classes.

The first two sections of Table 2 and Table 3 show the results for our zero-shot experimentation using LLMs for predicting the political bias on a 3- and 5-point scale, respectively.

In the 3-point setting, LLaMA3-70B achieves the highest accuracy and recall when only the media name is used, highlighting its stronger knowledge. Interestingly, smaller models such as LLaMA2-7B and Mistral-7B achieve a better MAE, indicating less confusion between the classes, in addition to their subpar accuracy compared to LLaMA3-70B. We observe the same trend for political bias in the 5-point setting.

However, when articles from the respective media are added to the prompt, the accuracy of the models for both settings decreases, contrary to the trend observed while predicting the factuality of reporting. This suggests that judging the political bias of a media outlet based on a single article at a time is a difficult task for LLMs. This highlights the need for an approach that assesses media bias on fine-grained topics before determining the overall political bias, as demonstrated in our methodology, which yielded better results.

5.5 Impact of Media Popularity

We conducted an error analysis to analyze the relationship between media popularity and model performance. The objective of this analysis was twofold: *(i)* to determine whether the labeling performance of the LLM-based approach correlates with the popularity of the media outlets; and *(ii)* to identify systematic challenges in classifying the less popular or newer outlets.

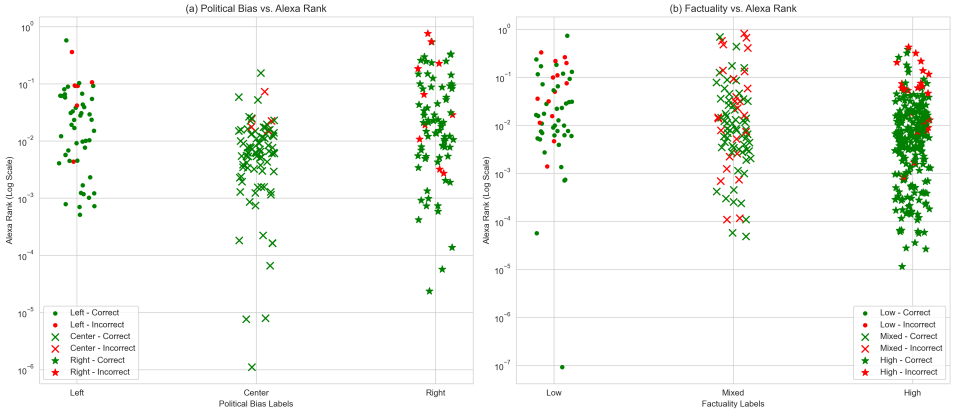

Figure 2: Best model performance vs. media outlet popularity. (a) Political bias labels, and (b) Factuality labels plotted against Alexa Rank (log scale). Each point represents a media outlet with its original label. Green markers indicate correct predictions, and red markers indicate errors. A lower Alexa Rank means a more popular medium.

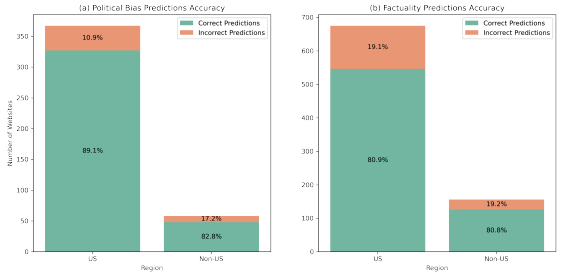

Figure 3: Correct vs. incorrect predictions for U.S. and non-U.S. media outlets, highlighting higher accuracy for U.S.-based outlets.

We used the Alexa Rank feature, which measures site popularity, as provided by Panayotov et al. (2022). We used the subset of data from our test set for which the Alexa Rank metric was available, as the Alexa Rank service is no longer live. This subset was analyzed to evaluate the ratio of media outlets correctly labeled by our best-performing models for both political bias and factuality of the reporting. We plotted the popularity of each media outlet on a logarithmic scale against its corresponding label. A lower Alexa Rank indicates a more popular media outlet.

Figure 2 shows a plot of the media outlets according to their original labels. Notably, a red cluster appears towards the top of the figure, indicating that the model struggles to predict the correct labels for less popular media outlets. Conversely, we observe more green clusters towards the bottom, which indicates that the model performs better at labeling more popular outlets. This pattern suggests that the model benefits from prior knowledge likely encoded in the LLM for well-known outlets.

Ablation: Political Bias Prediction		
Data Configuration	**Acc.↑**	**MAE↓**
Leaning	0.869	0.144
Reason	0.905	0.106
Leaning + Reason	**0.937**	**0.075**

Table 4: Ablation study on political bias prediction using different data configurations.

We extended our analysis to compare model performance on U.S. vs. non-U.S. media outlets, addressing potential regional bias in the LLM's training. Figure 3 shows that the model performs better on U.S.-based outlets, supporting the assumption that it has greater exposure to U.S. sources during training.

Overall, these findings show that LLMs label popular and U.S.-based media outlets more accurately but struggle with less popular, newer, or non-U.S. outlets. This highlights the need for improved methods to classify emerging or less popular outlets and to mitigate regional bias for better performance across diverse media.

5.6 Ablation Study

We conducted an analysis of the impact of different data configurations on the political bias prediction task using our best-performing model. As shown in Table 4, we experimented with three training setups: *(i)* using only the leaning information from GPT responses, resulting in 86.90% accuracy; *(ii)* using only the reasons from GPT responses, achieving 90.50% accuracy; and *(iii)* using both leaning and reasoning, yielding 93.50% accuracy.

These results highlight that LLM reasoning provides critical information and that combining both leaning and reasoning leads to better performance, demonstrating the importance of using multiple types of data for improved accuracy.

6 Conclusion & Future Work

We presented a comprehensive study on detecting political bias and factuality in news media. We collected data from LLMs using handcrafted prompts and a systematic method integrating professional fact-checking criteria. The models trained on these data demonstrated sizeable improvements over the existing approaches. Moreover, our experiments also revealed that without these expert guidelines in the prompts, most LLMs struggle to accurately classify the political bias and the factuality of the reporting of the news media. This underscores the crucial role of expert guidelines in improving the reliability and accuracy of LLM-based assessments.

Our methodology achieves 80.60% accuracy and an MAE of 0.206 for factuality prediction, while 93.50% accuracy and an MAE of 0.075 for political bias (3-point scale) prediction with 4,192 and 2,142 labeled media outlets, respectively. Previous work by Baly et al. (2019) achieved the best MAE of 0.481 for factuality prediction and 1.475 for political bias prediction with 949 labeled media outlets. Later work by Baly et al. (2020a) achieved 71.52% accuracy for factuality and 85.29% for political bias prediction. More recent work by Panayotov et al. (2022) using the same dataset, achieved 74.27% and 92.08% accuracy for factuality and political bias tasks, respectively. Our results compare favorably to these previous works, especially given our larger and more diverse labeled dataset.

In future work, we plan to refine factuality assessment by incorporating expert methodologies directly into prompts, expand political bias detection beyond U.S.-centric labels, and jointly predict factuality and political bias for a more comprehensive assessment. We will explore prompt learning and optimization techniques like APO (Pryzant et al., 2023), to reduce prompt bias. We will try retrieval augmentation and graphical features (related sites) for improved robustness. While cost considerations limited our exploration of GPT-4 variants and fine-tuning approaches, we plan to address these in the future. Finally, we aim to use open-source, instruction-tuned models beyond OpenAI's GPT-3.5 for reproducibility and community support.

Limitations

One limitation of our work is the use of the *GPT-3.5-turbo-1106* for data curation and interpretation, which may be influenced by factors such as data quality and diversity. Additionally, while the dataset includes a broad range of news outlets, it is primarily sourced from the MBFC, which is a U.S.-centric point of view database, and English-language outlets. This limits the model's generalizability to media from other regions.

Our methodology involves querying LLMs with specific prompts, which may lead to biased or incomplete assessments of less familiar media outlets due to biases such as training data bias, cultural bias, and confirmation bias. These biases can potentially skew evaluations by reflecting predominant viewpoints in the LLM's training data, affecting the model's objectivity. While our approach shows promising results, a key limitation arises when assessing media outlets not encountered during training. The model's generalizability to such outlets remains uncertain, highlighting the need for further investigation and additional strategies to ensure accurate assessments across a broader range of sources.

Moreover, we rely on methodologies that could be further refined by incorporating expert criteria, and our political bias detection approach is predominantly U.S.-centric. Extending beyond the left/center/right labels to capture a more nuanced political spectrum could yield more comprehensive insights. Furthermore, we have not fully explored the joint prediction of factual reporting levels and political bias, which could enhance our overall assessment. Potential failure modes, such as misclassification due to ambiguous language, incorrect factuality assessments from an insufficient context, and hallucinations from LLM responses, require attention to improve the robustness and reliability of our methodology. Additionally, integrating graphical features and leveraging retrieval augmentation to provide external evidence remain areas for improvement. Due to cost constraints, we have not experimented with GPT-4 variants, and while prompting has been more practical thus far, future work will consider fine-tuning approaches. Finally, we plan to adopt more open-source instruction-tuned models beyond OpenAI's GPT-3.5 to bolster reproducibility and mitigate potential model discontinuations. We acknowledge these limitations and intend to address them in our future research.

We acknowledge hallucinations in LLMs are inevitable (Xu et al., 2024), requiring validation through cross-referencing with external databases and expert reviews. Other limitations arise from limited labeled data, potential biases, and challenges in capturing nuanced political bias and factuality. Further research using diverse human-generated datasets beyond MBFC, incorporating human-in-the-loop approaches (Klie et al., 2020), and exploring different models is crucial to enhance the robustness and applicability of our findings.

Ethical Statement

We made every effort to ensure that the analysis and interpretation of the data were conducted impartially and objectively, avoiding any undue bias or prejudice. Transparency in the reporting methodologies and findings was emphasized to facilitate open dialogue and critical discussion within the academic community and beyond. An important ethical consideration revolved around the potential impact of the research findings on various stakeholders, including media outlets, journalists, and the general public. The findings could influence public trust in the media, highlight systemic biases, or shape how media outlets approach content creation and labeling. While this line of research holds promise in promoting media accountability and transparency, it also carries risks, such as unfairly stigmatizing certain outlets or reinforcing existing biases when misinterpreted. Careful communication of these results is crucial to prevent misuse or misrepresentation. To protect the integrity and privacy of the sources, any articles from the news media used during the analysis are not to be released; only the scraping recipes are to be shared.

Bias

We recognize that our methodology may be susceptible to biases, including misclassification from ambiguous language, errors due to limited context from summarization, and hallucinations from LLM responses. We also acknowledge potential prompt bias influencing model outputs. While we have taken serious measures to mitigate these issues, they persist as challenges. To systematically address them, we plan to explore prompt learning and optimization techniques such as APO (Pryzant et al., 2023). These steps aim to enhance the robustness and reliability of our approach as we improve our methodology in future work.

References

Dilshod Azizov, Zain Muhammad Mujahid, Hilal AlQuabeh, Preslav Nakov, and Shangsong Liang. 2024. SAFARI: cross-lingual bias and factuality detection in news media and news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 12217–12231. Association for Computational Linguistics.

Dilshod Azizov, Preslav Nakov, and Shangsong Liang. 2023. Frank at checkthat!-2023: Detecting the political bias of news articles and news media. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 289–305. CEUR-WS.org.

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 EMNLP*, Brussels, Belgium. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James R. Glass, and Preslav Nakov. 2020a. 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, ACL 2020, Online, July 5-10, 2020*, pages 3364–3374. 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.

Ramy Baly, Giovanni Da San Martino, James R. 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 2020, Online, November 16-20, 2020*, pages 4982–4991. 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. 2023. The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *45th ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*. 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, Maram Hasanain, Mücahid Kutlu, Chengkai Li, Federico Ruggeri, Julia Maria Struß, and Wajdi Zaghouani. 2023. Overview of the CLEF-2023 checkthat! lab on checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF*, Lecture Notes in Computer Science.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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*, Thessaloniki, Greece.

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)*, Minneapolis, Minnesota. Association for Computational Linguistics.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2016. Knowledge-based trust: Estimating the trustworthiness of web sources. *IEEE Data Eng. Bull.*, 39(2):106–117.

Matthew Gentzkow and Jesse M. Shapiro. 2006. Media bias and reputation. *Journal of Political Economy*, 114(2):280–316.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 235–238, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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*.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From Zero to Hero: Human-In-The-Loop Entity Linking in Low Resource Domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6982–6993. 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*, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Muhammad Arslan Manzoor, Ruihong Zeng, Dilshod Azizov, Preslav Nakov, and Shangsong Liang. 2025. MGM: global understanding of audience overlap graphs for predicting the factuality and the bias of news media. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 7279–7295. Association for Computational Linguistics.

Nikhil Mehta and Dan Goldwasser. 2024. An interactive framework for profiling news media sources. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 40–58. Association for Computational Linguistics.

Salman Bin Naeem, Rubina Bhatti, and Aqsa Khan. 2021. An exploration of how fake news is taking over social media and putting public health at risk. *Health Information & Libraries Journal*, 38(2):143–149.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray,

John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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 EMNLP*, pages 7470–7480, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gordon Pennycook and David G. Rand. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25(5):388–402.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 295th ACM International Conference on Information & Knowledge Management*, CIKM '16, page 2173–2178, New York, NY, USA. Association for Computing Machinery.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 155–158. ACM.

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.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7957–7968. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on EMNLP*, Singapore. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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 EACL*, St. Julians, Malta. Association for Computational Linguistics.

Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. 2019. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, Minneapolis, Minnesota, USA. 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*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024.

Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Kai-Cheng Yang and Filippo Menczer. 2025. Accuracy and political bias of news source credibility ratings by large language models. In *Proceedings of the 17th ACM Web Science Conference 2025, Websci 2025, New Brunswick, NJ, USA, May 20-24, 2025*, pages 127–137. ACM.

Appendix

A Handcrafted Prompts

A.1 Stance on Public Figures or Topics

A sample response for the prompt used in § 3.1.1 can be seen in Listing 1.

A.2 Stance on Current Popular Topics

A sample response for the prompt used in § 3.1.2 can be seen in Listing 2.

A.3 Determining Factuality of Reporting

A sample response for the prompt used in § 3.1.3 can be seen in Listing 3.

B Systematic Prompts

B.1 Systematic Guidelines

Left- and right-winged definitions of the 16 topics as described in § 3.2 can be seen in Listing 4.

B.2 Response to Systematic Prompt

A sample response for the prompt used in § 3.2 can be seen in Listing 5.

C Zero-Shot Predictions

C.1 LLM Response Given a Media Name

A response where LLM was unable to predict the factuality of reporting:

```
{
    "input": "vancouvertimes.org",
    "output": -1
}
```

C.2 Political Bias Prompt Given a Media Name

The following is the prompt used to predict the political bias of the media given its name:

```
system_prompt = '''You are an expert in media
↪  analysis. Classify the bias of a media given
↪  its name ONLY from one of three categories
↪  from the list provided below:
- left
- center
- right
Return -1 if you can not classify.'''
user_prompt = "{domain}"
```

MBFC Dataset			
Political Bias		Factuality	
Left	398	Low	597
Left-Center	600	Mixed	1200
Center	913	High	2395
Right-Center	907		
Right	831		
Total	**3649**	**4192**	

Table 5: Distribution of labels in our dataset.

C.3 Summarization

The following prompt in Listing 6 is used to summarize the articles from the news media as described in § 4.2, where the **{article}** placeholder is replaced with the article from the news media.

C.4 Political Bias Prompt Given Media Articles

The following is the prompt used to predict the political bias of the media given its summarized articles:

```
system_prompt = '''You are an expert in media
↪  analysis. Classify the bias of the given news
↪  article of {media} into one of three
↪  categories from the list provided below:
- left
- center
- right
Return -1 if you can not classify.'''
user_prompt = "{article}"
```

D Dataset

An example annotation for cnn.com provided by MBFC is shown in Figure 4. Detailed dataset statistics are shown in Table 5

```
{
    "foxnews.com": {
        "Trump": {
            "stance": "pro-Trump",
            "reason": "Fox News has been known for its conservative-leaning coverage and has been
            ↪  generally supportive of President Trump's policies and actions."
        },
        "Biden": {
            "stance": "anti-Biden",
            "reason": "Fox News is generally known for its conservative and right-leaning viewpoints,
            ↪  and it has often been critical of President Biden and his policies."
        },
        "Putin": {
            "stance": "Fox News is generally critical of Putin.",
            "reason": "Fox News has often reported on Putin's controversial actions and policies, such
            ↪  as his involvement in the conflict in Ukraine and his crackdown on political
            ↪  opposition within Russia."
        },
        "American": {
            "stance": "Fox News is generally considered to have a conservative bias, and it often
            ↪  promotes pro-American viewpoints.",
            "reason": "The network's programming and commentary often align with conservative and
            ↪  nationalist perspectives, which are often associated with pro-American sentiments."
        },
        "Russian": {
            "stance": "Fox News is not generally considered to be pro-Russian. It is known for its
            ↪  conservative and right-leaning perspective on news and politics.",
            "reason": "Fox News has been critical of Russia and its leadership on various occasions,
            ↪  particularly in relation to foreign policy and national security issues."
        },
        "German": {
            "stance": "Neutral",
            "reason": "Fox News does not have a specific stance on Germany as a whole. Its coverage of
            ↪  Germany may vary depending on the specific news story or issue being discussed."
        },
        "Palestinian": {
            "stance": "Fox News is generally perceived as having a pro-Israel stance, which may lead
            ↪  to a more critical view of the Palestinian perspective.",
            "reason": "Fox News has been known for its support of Israel and its policies, which may
            ↪  influence its coverage of the Palestinian issue."
        }
    }
}
```


Listing 1: Sample response for the prompt used in § 3.1.1.

```
{
    "foxnews.com": {
        "Ukraine": {
            "stance": "Fox News generally takes a conservative stance on Ukraine, often highlighting
            ↪  the threat posed by Russia and advocating for a strong U.S. response.",
            "reason": "Fox News tends to align with conservative viewpoints and policies, which often
            ↪  include a more hawkish approach to foreign policy and national security."
        },
        "Climate Change": {
            "stance": "Skeptical",
            "reason": "Fox News has been known to feature content that questions the scientific
            ↪  consensus on climate change and often provides a platform for climate change sceptics
            ↪  and critics."
        },
        "Gun Control": {
            "stance": "Fox News generally opposes strict gun control measures, arguing for the
            ↪  protection of Second Amendment rights and advocating for a focus on mental health and
            ↪  law enforcement efforts to address gun violence.",
            "reason": "Fox News often presents arguments in favour of gun rights and against gun
            ↪  control legislation, emphasizing the importance of individual freedoms and the
            ↪  potential ineffectiveness of stricter gun laws in preventing crime."
        },
        "Immigration": {
            "stance": "Fox News generally takes a conservative stance on immigration, advocating for
            ↪  stricter border control and enforcement of immigration laws.",
            "reason": "Fox News often emphasizes the need for national security and the rule of law,
            ↪  and highlights the potential economic and social impacts of immigration."
        },
        "Abortion": {
            "stance": "Against",
            "reason": "Fox News generally takes a conservative stance on social issues, including
            ↪  abortion, and often features content that aligns with pro-life viewpoints."
        }
    }
}
```

Listing 2: Sample response for the prompt used in § 3.1.2.

```
{
    "foxnews.com": {
        "Q1": {
            "answer": "No",
           "reason": "Fox News has been criticized for biased reporting and spreading misinformation."
        },
        "Q2": {
            "answer": "Yes",
            "reason": "Fox News has failed fact-checks by multiple fact-checking organizations."
        },
        "Q3": {
            "answer": "Right",
            "reason": "Fox News is known for its conservative and right-leaning political stance."
        },
        "Q4": {
            "answer": "Yes",
           "reason": "Fox News is known for having a conservative bias in its reporting and editorial
            ↪  content."
        },
        "Q5": {
            "answer": "Conservative, News, Opinion",
            "reason": "Fox News is known for its conservative political stance and provides news and
            ↪  opinion content."
        },
        "Q6": {
            "answer": "Conservative, Right-leaning, Partisan",
            "reason": "Fox News is known for its conservative and right-leaning editorial stance, and
            ↪  it is often considered to have a partisan bias in its reporting."
        }
    }
}
```


Listing 3: Sample response for the prompt used in § 3.1.3.

```
{
    "General Philosophy":{
        "left": "Collectivism: Community over the individual. Equality, environmental protection,
        ↪  expanded educational opportunities, social safety nets for those who need them.",
        "right": "Individualism: Individual over the community. Limited Government with Individual
        ↪  freedom and personal property rights. Competition.",
    },
    "Abortion":{
        "left": "Legal in most cases.",
        "right": "Generally illegal with some exceptions.",
    },
    "Economic Policy":{
        "left": "Income equality; higher tax rates on the wealthy; government spending on social
        ↪  programs and infrastructure; stronger regulations on business. Minimum wages and some
        ↪  redistribution of wealth.",
        "right": "Lower taxes; less regulation on businesses; reduced government spending.  The
        ↪  government should tax less and spend less. Charity over social safety nets. Wages should
        ↪  be set by the free market.",
    },
    "Education Policy":{
        "left": "Favor expanded free, public education. Reduced cost or free college.",
        "right": "Supports homeschooling and private schools. Generally not opposed to public
        ↪  education, but critical of what is taught.",
    },
    "Environmental Policy":{
        "left": "Regulations to protect the environment. Climate change is human-influenced and
        ↪  immediate action is needed to slow it.",
        "right": "Considers the economic impact of environmental regulation. Believe the free market
        ↪  will find its own solution to environmental problems, including climate change. Some deny
        ↪  climate change is human-influenced.",
    },
    "Gay Rights":{
        "left": "Generally support gay marriage; support anti-discrimination laws to protect LGBT
        ↪  against workplace discrimination.",
        "right": "Generally opposed to gay marriage; opposed to certain anti-discrimination laws
        ↪  because they believe such laws conflict with certain religious beliefs and restrict
        ↪  freedom of religion.",
    },
    "Gun Rights":{
        "left": "Favors laws such as background checks or waiting periods before buying a gun; banning
        ↪  certain high capacity weapons to prevent mass shootings.",
        "right": "Strong supporters of the Second Amendment (the right to bear arms), believing it's a
        ↪  deterrent against authoritarian rule and the right to protect oneself. Generally, does
        ↪  not support banning any type of weaponry.",
    },
    "Health Care":{
        "left": "Most support universal healthcare; strong support of government involvement in
        ↪  healthcare, including Medicare and Medicaid. Generally, support the Affordable Care Act.
        ↪  Many believe healthcare is a human right.",
        "right": "Believe private companies can provide healthcare services more efficiently than
        ↪  government-run programs. Oppose the Affordable Care Act. Insurance companies can choose
        ↪  what to cover and compete with each other. Healthcare is not a right.",
    },
```

```
    {
    "Immigration":{
        "left": "Generally, support a moratorium on deporting or offering a pathway to citizenship to
        ↪  certain undocumented immigrants. e.g., those with no criminal record have lived in the
        ↪  U.S. for 5+ years. Less restrictive legal immigration.",
        "right": "Generally against amnesty for any undocumented immigrants. Oppose a moratorium on
        ↪  deporting certain workers. Funding for stronger enforcement actions at the border
        ↪  (security, wall). More restrictive legal immigration.",
    },
    "Military":{
        "left": "Decreased Spending",
        "right": "Increased Spending",
    },
    "Personal Responsibility":{
        "left": "Strong government to provide a structure. Laws are enacted to protect every
        ↪  individual for an equal society. Safety nets for those in need.",
        "right": "Personal responsibility and it is the government's role to hold them accountable.
        ↪  Fair competition over safety nets.",
    },
    "Regulation":{
        "left": "Government regulations are needed to protect consumers and the environment.",
        "right": "Government regulations hinder free-market capitalism and job growth.",
    },
    "Social Views":{
        "left": "Based on community and social responsibility. Gay couples to get equal rights like
        ↪  everyone else (e.g. marriage); abortion should be legal; support embryonic stem cell
        ↪  research. Support restrictions and regulations around the right to bear arms. Supports
        ↪  taxpayer funding of Planned Parenthood.",
        "right": "Based on individual rights and justice. Opposed to gay marriage, abortion, and
        ↪  embryonic stem cell research. Support the right to bear arms, death penalty, and personal
        ↪  responsibility as an individual. Opposed to taxpayer funding of Planned Parenthood.",
    },
    "Taxes":{
        "left": "Progressive taxation (high-income earners taxed at a higher rate). Generally not
        ↪  opposed to raising taxes to fund the government.",
        "right": "Tend to favor a "flat tax" (same tax rate regardless of income). Generally opposed
        ↪  to raising taxes to fund the government.",
    },
    "Voter ID":{
        "left": "Against voter ID laws citing an undue burden on lower-income groups causing them to
        ↪  be disenfranchised, and that there is virtually no evidence of voter fraud actually
        ↪  occurring.",
        "right": "For voter identification laws to combat alleged voter fraud.",
    },
    "Business Rights":{
        "left": "Supports unions and worker protections. Raising the minimum wage. Higher corporate
        ↪  taxes.",
        "right": "Favors business owners and corporations with the expectation higher profits will
        ↪  result in higher wages through a free-market. Generally opposed to a minimum wage. Lower
        ↪  corporate taxes.",
    }
}
```

Listing 4: Left- and right-winged definitions of the 16 topics as described in § 3.2.

```
{
    "foxnews.com": {
        "General Philosophy": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on
            ↪  political and social issues, promoting individualism and limited government."
        },
        "Abortion": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on social
            ↪  issues, including abortion. It generally aligns with the right-leaning stance of
            ↪  being generally illegal with some exceptions."
        },
        "Economic Policy": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on
            ↪  economic policy, advocating for lower taxes, less regulation on businesses, and
            ↪  reduced government spending."
        },
        "Education Policy": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective, often
            ↪  supporting homeschooling and private schools while being critical of public education
            ↪  policies."
        },
        "Environmental Policy": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on
            ↪  environmental policy, often emphasizing the economic impact of regulations and
            ↪  expressing skepticism about human-influenced climate change."
        },
        "Gay Rights": {
            "leaning": "right",
            "reason": "Fox News is generally opposed to gay marriage and may be opposed to certain
            ↪  anti-discrimination laws due to religious beliefs."
        },
        "Gun Rights": {
            "leaning": "right",
           "reason": "Fox News is known for its conservative and right-leaning perspective on various
            ↪  issues, including gun rights. They generally support the Second Amendment and the
            ↪  right to bear arms without significant restrictions."
        },
        "Health Care": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on
            ↪  healthcare, often opposing government involvement and supporting private companies in
            ↪  providing healthcare services."
        },
        "Immigration": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on
            ↪  immigration, often advocating for stronger enforcement actions at the border and
            ↪  opposing amnesty for undocumented immigrants."
        },
        "Military": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective, often
            ↪  advocating for increased military spending and a strong military presence."
        },
```

```
    {
        "Personal Responsibility": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective, often
            ↪  emphasizing personal responsibility and limited government intervention."
        },
        "Regulation": {
            "leaning": "right",
           "reason": "Fox News is known for its conservative and right-leaning perspective, which often
            ↪  opposes government regulations as hindering free-market capitalism and job growth."
        },
        "Social Views": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective on social
            ↪  issues, including opposition to gay marriage, abortion, and embryonic stem cell
            ↪  research, as well as support for the right to bear arms and opposition to taxpayer
            ↪  funding of Planned Parenthood."
        },
        "Taxes": {
            "leaning": "right",
            "reason": "Fox News tends to favor a 'flat tax' and is generally opposed to raising taxes
            ↪  to fund the government."
        },
        "Voter ID": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective, and it
            ↪  generally supports voter identification laws to combat alleged voter fraud."
        },
        "Business Rights": {
            "leaning": "right",
            "reason": "Fox News is known for its conservative and right-leaning perspective, often
            ↪  favoring business owners and corporations with the expectation of higher profits
            ↪  resulting in higher wages through a free-market approach."
        }
    }
}
```

Listing 5: A sample response for the prompt used in § 3.2.

```
system_prompt = '''Summarize the following news article in 250-300 words. Ensure the summary covers
↪  the article's key points and main details.'''
user_prompt = "{article}"
```

Listing 6: Prompt used to summarize the articles from news media as described in § 4.2.

CNN – Bias and Credibility

Factual Reporting
Very High
High
MOSTLY FACTUAL
Mixed
Low
Very Low

LEFT-CENTER BIAS

These media sources have a slight to moderate liberal bias. They often publish factual information that utilizes loaded words (wording that attempts to influence an audience by appeals to emotion or stereotypes) to favor liberal causes. These sources are generally trustworthy for information but may require further investigation. See all Left-Center sources.

- **Overall, we rate CNN moderately left-center biased based on editorial positions by TV hosts that consistently favor the left, while straight news reporting falls just left of center through bias by omission. We also rate them as Mostly Factual in reporting rather than high due to two failed fact checks in the last five years.**

Detailed Report

Bias Rating: LEFT-CENTER
Factual Reporting: MOSTLY FACTUAL

Figure 4: An example of annotation of a news outlet from MBFC. Source: www.mediabiasfactcheck.com.