A Multi-View Media Profiling Suite: Resources, Evaluation, and Analysis

Muhammad Arslan Manzoor[α,β*] **Dilshod Azizov**[α*] **Daniil Orel**[α*]
Umer Siddique[γ] **Zain Muhammad Mujahid**[δ] **Yufang Hou**[β] **Preslav Nakov**[α]

[α]MBZUAI, UAE [β]Interdisciplinary Transformation University, Austria
[γ]University of Texas at San Antonio, USA [δ]University of Copenhagen, Denmark
{muhammad.arslan, preslav.nakov}@mbzuai.ac.ae

Abstract

News outlets shape public opinion on a scale that makes automated detection of political bias and factuality essential. Yet, the field still lacks unified resources, comprehensive evaluations in diverse approaches, and systematic analyzes of the representations and fusion strategies that matter the most, especially under label sparsity and dataset diversity. In addition, there is little empirical work that reports broad observation-driven findings about what consistently works, what fails, and why. We address these gaps with four contributions: *(i)* **MBFC-2025**, a large-scale label set that covers ∼2,600 outlets from Media Bias/Fact Check (MBFC); *(ii)* multi-view representations for ACL-2020 (Panayotov et al., 2022) (∼900 outlets) and MBFC-2025, spanning Alexa graphs, hyperlink graphs, LLM-derived graphs, articles, and Wikipedia descriptions; *(iii)* systematic evaluation and analysis of embedding views and fusion strategies, including an RL-based fusion variant; and *(iv)* extensive experiments that achieve state-of-the-art results on ACL-2020 and establish strong benchmarks on MBFC-2025.

1 Introduction

Profiling of news outlets has become increasingly important (Baly et al., 2020b; Panayotov et al., 2022; Mehta and Goldwasser, 2023; Mehta et al., 2022) because claim- or article-level verification is costly and difficult to scale. Outlet-level profiling offers a practical alternative: by characterizing sources through their historical *political bias* and *factuality*, it enables scalable, real-time monitoring. A widely used source of supervision is Media Bias/Fact Check (MBFC), which provides expert outlet labels and has served as the ground truth in many studies (Mehta and Goldwasser, 2023; Panayotov et al., 2022; Baly et al., 2018, 2019; Hounsel et al., 2020; Nakov et al., 2024). However, effective profiling depends on reliable outlet representations.

Most existing methods rely on textual content (Baly et al., 2020a, 2019), which can be noisy (frame, boilerplate), difficult to crawl, and expensive to collect on scale. Thus, recent work explores alternative signals, including infrastructure and website metadata (Hounsel et al., 2020), social context such as followers (Baly et al., 2020b), and graph-based representations that connect outlets through audiences or hyperlinks (Panayotov et al., 2022; Manzoor et al., 2025). While promising, graph-based approaches can suffer from structural sparsity (*e.g.,* disconnected components) that limits representation learning (Longa et al., 2024; Zhang et al., 2024), and prior evaluations are often tied to a single graph view and benchmark, leaving open which signals and fusion strategies are most reliable under missing or noisy modalities.

To study these questions, we build a *multi-view* profiling suite that combines three graph views and two textual views. The graph views include: *(i) Alexa graphs* capturing audience overlap, *(ii) hyperlink graphs* reflecting inter-outlet links, and *(iii) LLM-derived graphs* encoding semantic similarity between outlets. The textual views include: *(iv) outlet articles* capturing tone and framing, and *(v) Wikipedia descriptions* providing context. We evaluate these views in the ACL-2020 (Panayotov et al., 2022) label set, which covers fewer than 900 outlets with 3-point political bias and factuality scales, and where many outlets lack article- or Wikipedia-based representations due to missing content and crawling constraints. To improve coverage and granularity, we introduce **MBFC-2025**, a source-level benchmark of ∼2,6k outlets with 5-point scale labels for both tasks. We construct the same multi-view representations for MBFC-2025 and evaluate embedding and fusion strategies to identify effective combinations. We also include a reinforcement learning (RL)-based fusion variant, formulated as a contextual bandit, to explore dynamic weighting under noisy and incomplete views.

[*]These authors contributed equally to this work.

Findings of the Association for Computational Linguistics: ACL 2026, pages 25265–25286

We train graph neural networks (GNNs) over the media graphs and fine-tune pretrained language models (PLMs) on articles and Wikipedia entries. Our empirical study provides strong baselines, clarifies which representations contribute the most, and yields state-of-the-art results on ACL-2020 while establishing new benchmarks on MBFC-2025.

Our contributions are as follows:

- **Benchmark.** We introduce **MBFC-2025**, a large-scale source-level benchmark annotated by MBFC with ∼2,600 outlets labeled for political bias and factuality on a 5-point scale.
- **Resources.** We compile five-view representations for ACL-2020 and MBFC-2025: three graphs and two text representations.
- **Evaluation and Analysis.** We systematically evaluate embedding views and fusion strategies (including an RL-based fusion variant), provide empirical analyzes of what works and why, and achieve state-of-the-art results on ACL-2020 while establishing strong benchmarks on MBFC-2025*.

2 Related Work

Political bias: Political bias reflects the ideological orientation of text that influences opinions and voting behavior (Druckman and Parkin, 2005; Boyle et al., 2007; Prior, 2013). Early work focused on textual analysis (Afroz et al., 2012; Pérez-Rosas et al., 2018; Conroy et al., 2015; Da San Martino et al., 2023; Barrón-Cedeño et al., 2023a,b; Azizov et al., 2023), while later studies incorporated contextual signals, including multimedia, infrastructure, and social context (Baly et al., 2020b; Hounsel et al., 2020; Castelo et al., 2019; Fairbanks et al., 2018). Recent work uses PLMs and multimodal data, including BERT-based models (Guo et al., 2022), text-image methods (Qiu et al., 2022), and ideology-driven pretraining (Liu et al., 2022). More recently, prompt-based methods (Maab et al., 2024; Mujahid et al., 2025), adaptable representations (Lin et al., 2024), and cross-lingual techniques (Azizov et al., 2024) have been explored.

The reliability of the media is commonly assessed by aligning the predictions of the model with the expert ratings (Yang and Menczer, 2025). Earlier approaches relied on positions towards claims (Mukherjee and Weikum, 2015; Popat et al., 2017), while later work used gold labels and PLMs for outlet profiling (Baly et al., 2020b).

Mehta and Goldwasser (2023) combine graph models, PLMs, and human input for fake news detection. RL has been explored for reliability prediction (Burdisso et al., 2024), but primarily for scoring rather than representation fusion.

GNNs: GNNs are used for representation learning in media analysis (Mehta et al., 2022; Manzoor et al., 2025). Panayotov et al. (2022) construct homophilic graphs showing that similar outlets attract audiences. Mehta and Goldwasser (2023) integrate graph models, LLMs, and human input without requiring labels. Other work incorporates multihop reasoning (Zhang et al., 2022) and event graphs (Lei and Huang, 2024) to improve bias detection.

Fusion: Media graphs suffer from sparsity and few labels, hindering GNNs from capturing dependencies (Cui et al., 2020; Sun et al., 2020; Bodnar et al., 2021). Previous work addresses this through multi-view learning, constructing graphs, and combining embeddings via strategies such as concatenation, averaging, or attention (Ding et al., 2022; Ma et al., 2020; Yuan et al., 2021). In contrast, we propose an RL-based fusion framework that dynamically learns outlet-specific weights over heterogeneous graph and textual embeddings. To our knowledge, this is the first application of RL for multi-view fusion in bias and factuality detection.

3 Data Construction Pipeline

We use MBFC as the label source for both datasets (details are in the Appendix A). ACL-2020 (Baly et al., 2020b) includes 859 media sources with 3-point labels for political bias and factuality (*left/center/right*, *high/mixed/low*), while MBFC-2025 uses a 5-point scale, adding *left-center*, *right-center*, *very low*, and *very high*. Next, we describe the representations and dataset statistics.

3.1 Graphs Construction

Generating diverse graphs and textual embeddings is key to address feature scarcity and enable data-driven models to better understand the underlying relationships and patterns.

In addition to the Alexa graph, we construct two new graphs: the Hyperlink graph and the LLM-graph, illustrated in Figure 1. They capture both inherent and implicit relationships between nodes. These relationships enable the generation of rich embeddings that ultimately complement other representations, *e.g.,* text-based embeddings, by addressing knowledge gaps.

*Our code is available at GitHub.

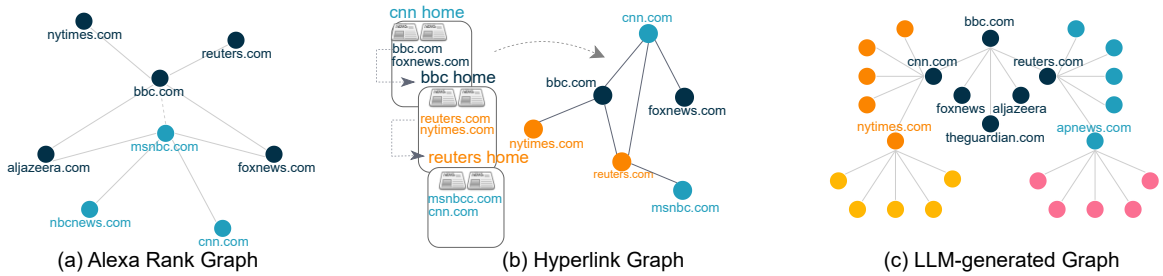


Figure 1: Illustration of generated graphs using (a) Alexa Rank tool (b) Hyperlink graph, and (c) LLM.

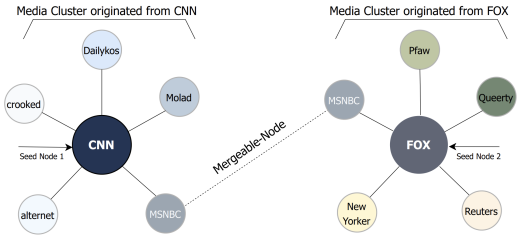


Figure 2: CNN and FOX are seed nodes in the labeled media set (ACL-2020 & MBFC-2025), while MSNBC serves as a mergeable node, forming a fundamental building block for graph construction.

Alexa Graph. Panayotov et al. (2022) used the Alexa Rank tool[†] to query the addresses of the news media, generating a list of 4-5 similar sites based on the overlap of the audience. We used these similar sites as nodes in a graph, with audience overlap defining the edges. Starting with websites from the ACL-2020 dataset as seed nodes, they iteratively expanded the graph by querying for each website and adding new nodes and edges. This iterative process was repeated five times, resulting in five levels of graphs, each progressively denser and more comprehensive.

Analysis of the constructed graphs in the Appendix B revealed several disconnected components, indicating isolated networks of nodes. As the graph levels increased, the number of disconnected components decreased, likely due to the growing number of nodes facilitating component merging.

Graph level 3, the highest publicly accessible level for both factuality and political bias tasks, was chosen to produce embeddings with GNNs. The Alexa Rank tool also provided node representations, treated as node attributes during GNN training. These representations, including site rank, total linked sites, bounce rate, and daily user time (Panayotov et al., 2022).

Hyperlink (On-site) Graph. To collect hyperlinks, we crawled with the newspaper3k[‡] parser and extracted external links. We scraped the main page and articles, collecting up to 50 articles per site. The nodes represent unique websites in the network, while edges are counted twice for each website pair, since every link from website A to B creates two edges: A→B and B→A. To address graph sparsity, we employed a layered parsing strategy that iteratively expands from dataset websites (level 0) to linked sites at subsequent levels, where level 0 includes only websites from the dataset, and each additional level incorporates websites referenced by those in the previous level.

LLM-Graph Techniques such as Chain-of-Thought (CoT) prompting (Wei et al., 2022), Least-to-Most prompting (Zhou et al., 2023), and Tree-of-Thought (ToT) (Yao et al., 2023), use LLM reasoning to tackle complex tasks. However, to leverage the contextual knowledge of LLMs for this task, we prompt the model to retrieve five websites similar to a given media outlet x. The intuition behind this approach is two-fold: First, rather than directly asking the LLM to profile a media source, we leverage its contextual understanding to generate smaller clusters, seeded from ACL-2020 and MBFC-2025 datasets, which contain media labeled for factuality and political bias.

These clusters serve as the foundation for constructing media graphs, where we identify and merge the same nodes as shown in Figure 2. Second, this approach maintains consistency with other graph construction methods, where the number of similar nodes retrieved n is limited to five or fewer. To model the relationship between media outlets, we use GNNs, which generate rich media representations that are aware of the broader media landscape, including connections and distances between different media.

[†] http://www.alexa.com/siteinfo

[‡] https://github.com/codelucas/newspaper

Political Bias	Left	Left-center	Center	Right-center	Right	Total
ACL-2020	152	-	178	-	145	475
MBFC-2025	255	427	705	759	404	2550
Factuality	**Very High**	**High**	**Mixed**	**Low**	**Very Low**	
ACL-2020	-	295	119	61	-	475
MBFC-2025	30	1347	970	147	56	2550

Table 1: Label distribution statistics for political bias and factuality in the ACL-2020 (3-point scale) and MBFC-2025 (5-point scale) datasets.

To implement this, we used the OpenAI `Python` package to query the `API` endpoint of the *gpt-3.5-turbo-0125* model (GPT-3.5), released on January 25, 2024 (Ouyang et al., 2022; Brown et al., 2020). The specific prompt used is as follows:

```
prompt = "Based on similarity, give me 5 similar
↪  websites to, {domain}. Return only the
↪  websites URL, strictly without any
↪  explanation, don't add numbers in start. Each
↪  website should be wrapped under the tag <s>"
```

In this prompt, the **{domain}** placeholder is replaced with the domain of interest, such as *foxnews.com*. A sample response derived from GPT-3.5 for *bbc.com* is shown in Listing 1.

```
<s>https://www.cnn.com/</s>
<s>https://www.theguardian.com/</s>
<s>https://www.aljazeera.com/</s>
<s>https://www.nytimes.com/</s>
<s>https://www.reuters.com/</s>
```

Listing 1: Example of LLM response.

We process these responses to create a standard JSONL file for downstream analysis and storage. Next, we compile the level 0 graph, removing duplicates from websites that have already been processed to ensure data consistency. This procedure is repeated using our nested script to systematically generate graphs up to level 3. The final level 3 graph contains up to 18,508 nodes for ACL-2020 and 80,385 for MBFC-2025, as shown in Table 2.

In addition, our profiling emphasizes the input media rather than the URLs generated by the LLM, ensuring that the approach remains robust.

3.2 Textual Representations

To generate textual representations for media, we collect data from *Articles* and *Wikipedia* pages. For *Articles*, we extracted media details from MBFC, parsed news media pages with article links.

Dataset	Representation	# Sampled Label	# of Levels	# of Nodes	# of Edges
ACL-2020	Alexa graph	475	3	67,350	100,261
	LLM-graph	475	3	18,508	43,410
	Hyperlink graph	475	3	88,722	169,363
	Wikipedia	475	-	-	-
	Media (Articles)	475	-	-	-
MBFC-2025	Alexa graph	2,550	3	67,350	100,261
	LLM-graph	2,550	3	80,385	194,755
	Hyperlink graph	2,550	2	528,214	1,994,492
	Wikipedia	2,550	-	-	-
	Media (Articles)	2,550	-	-	-

Table 2: Statistics of representations, sampled labels, and graph structures for ACL-2020 and MBFC-2025.

Dataset	Task	Train	Dev	Test	Total
MBFC-2025	Political Bias & Factuality	2040	256	254	2550
ACL-2020	Political Bias & Factuality	383	43	49	475

Table 3: Statistics for train, development, and test sets.

Also, we retrieved article texts using custom scripts. Post-processing ensured that data were formatted into JSON. For *Wikipedia*, we gather media outlet information by extracting HTML from *Wikipedia* pages and converting it to JSON.

Furthermore, the goal of collecting *Articles* was *to capture what was written by the media outlet*, whereas *Wikipedia* pages were collected *to reflect what was written about the outlet*. Together, these sources were used to assess the media's current positioning, standing, or ideological leaning. Detailed data collection steps are provided in Appendix A.

3.3 Datasets Statistics

Table 1 presents the label distributions for political bias and factuality. Data were split into train/validation/test sets using an 80/10/10 ratio without overlap of media outlets or articles on the same story and event to prevent data leakage. Stratified sampling ensured balanced class distributions across splits. Table 2 summarizes the representation types, sampled labels, and graph structures, while Table 3 reports detailed statistics for each split in ACL-2020 and MBFC-2025 datasets.

4 Methodology

4.1 Problem Formulation

We formulate news media profiling as a multimodal (refers to the number of representations) classification problem, where political bias and factuality are predicted independently using various representations. Given a media outlet F_i, we represent it with a set of five representations:

$$\mathcal{F}_i = \{F^{(a)}, F^{(h)}, F^{(l)}\ F^{(t)}, F^{(w)}\}$$

Where $F^{(a)}$ denotes the Alexa graph, $F^{(h)}$ the Hyperlink graph, $F^{(l)}$ the LLM generated graph, $F^{(t)}$ the media articles, and $F^{(w)}$ the Wikipedia descriptions.

Let $\mathcal{G} : \mathcal{F} \rightarrow y$ denote a prediction function mapping the multi-view representations to a target label y, corresponds to either political bias or factuality. Our objective is to identify effective combinations of representations and fusion approaches that maximize predictive performance for each task.

4.2 Multi-View Integration

In this section, we present our multi-view integration framework, shown in Figure 3. For each target medium, we have a variety of representations, such as (*i*) *Alexa graph*, (*ii*) *Hyperlink graph*, (*iii*) *LLM-graph*, (*iv*) *Articles*, and (*v*) *Wikipedia*. Textual representations are extracted using advanced text models, while recent GNNs effectively handle graph representations. These are then fused using several complementary fusion strategies to provide a comprehensive view of the website's content and structure, and to find a better combination of representations.

4.2.1 GNN Models

In our study, we use three GNNs: Residual Gated Graph Convolutional Networks (ResGatedGCNs) (Bresson and Laurent, 2017), Graph Convolutional Networks (GraphConv) (Morris et al., 2019), and GraphSAGE (Hamilton et al., 2017). GraphConv extends convolutional operations to graphs, enabling the aggregation of information from neighboring nodes to learn node representations.

GraphSAGE introduces a sampling method that selects a fixed-size neighborhood for each node during training, reducing memory usage and computational costs while improving scalability for large graphs. ResGatedGCNs combine residual connections with gated mechanisms to mitigate over-smoothing, allowing for more stable information flow across layers.

4.2.2 Textual Models

We evaluate a diverse range of textual models, including: classical SVM with TF-IDF (Hearst et al., 1998; Sparck Jones, 1972), $\text{BERT}_{\text{Base}}$ (Devlin et al., 2019), $\text{RoBERTa}_{\text{Base}}$ (Liu et al., 2019), $\text{DistilBERT}_{\text{Base}}$ (Sanh et al., 2019), and $\text{ALBERT}_{\text{Base}}$ (Lan et al., 2020), Mistral 7B (Jiang et al., 2023), Llama-2 7B (Touvron et al., 2023), and GPT-4o (Achiam et al., 2023).

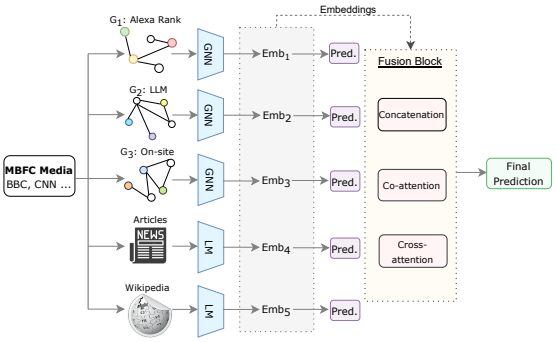


Figure 3: **End-to-end pipeline of our approach.** Given an MBFC media outlet, we construct multiple graphs and textual views. GNNs and pre-trained language models (PLMs) generate outlet-level embeddings for each view, followed by view-specific predictions. The embeddings are combined through various fusion mechanisms.

Textual representations: To obtain outlet-level labels, we aggregate individual article predictions using hard and soft voting. This ensemble approach (Freund and Schapire, 1997) helps reduce variance and ensures robust classification. We also used each outlet's Wikipedia description as an additional signal, capturing broader contextual and historical information. To leverage complementary information from both sources, we concatenated article and Wikipedia embeddings and applied the same aggregation strategies used for media articles.

4.2.3 Representations Fusion Approaches

We use $\text{DistilBERT}_{\text{Base}}$, chosen for efficiency, to generate embeddings from media articles and Wikipedia descriptions. To complement textual representations, we use GNNs to learn structural embeddings that capture relational patterns. We then combine textual and structural embeddings using fusion strategies. *(i) SVM Fusion:* Following Baly et al. (2020b), we concatenate embeddings and apply an SVM classifier. *(ii) MLP Fusion:* We apply a single-layer perceptron to combine embeddings, as done by Jaafar and Lachiri (2023), since this method outperforms simpler operations such as addition or multiplication. *(iii) Self-Attention Fusion:* We apply self-attention to the concatenated embeddings to capture feature importance. *(iv) Late Attention Fusion:* Namely, *(a) Cross-Attention*, where one representation attends to the other, and *(b) Co-Attention*, where both attend to each other. *(v) RL-based Dynamic Fusion:* To overcome static fusion limitations, we propose an RL-based method that learns outlet-specific weights for each view.

Hyper-parameter	ACL-2020				MBFC-2025			
	BERT	RoBERTa	DistilBERT	ALBERT	BERT	RoBERTa	DistilBERT	ALBERT
Batch size	80	95	110	120	80	90	100	100
Max length	256	256	512	512	128	128	256	256
Epochs	4	4	5	6	5	4	6	5
Learning rate	2e-5	2e-5	2e-5	2e-5	2e-5	2e-5	2e-5	2e-5

Table 4: Experimental setup for PLMs on ACL-2020 and MBFC-2025 datasets.

RL-based Dynamic Fusion. We cast fusion as a contextual bandit where actions do not affect state transitions. The state concatenates the multi-view embeddings for outlet F_i, $s_t = \{F^{(a)}, F^{(h)}, F^{(l)}, F^{(t)}, F^{(w)}\}$. The action is a continuous weight vector $w \in R^5$ with $w_k \in [0, 1]$ indicating the importance of each view. We compute the fused embedding as $E_{\text{fused}} = \sum_{k=1}^{5} w_k F^{(k)}$, and define the reward as the probability of the true-label under a fixed classifier, $r_t = P(y_{\text{true}} \mid E_{\text{fused}})$. The goal is to learn a policy $\pi(a \mid s)$ that maximizes the immediate expected reward. We implement the environment in Gymnasium (Towers et al., 2024) and train the policy with Proximal Policy Optimization (PPO) (Schulman et al., 2017) in Stable-Baselines3 (Raffin et al., 2021), setting the discount factor near zero to emphasize immediate rewards.

5 Experiments and Evaluation

5.1 Experimental Setup

GNNs. We learn media graph embeddings using a contrastive objective in an unsupervised setting, where labeled nodes constitute under 1%. For each graph, we obtain 64-dimensional embeddings using GCN, GraphSAGE, and ResGatedGCN. All models share the same hyperparameters: *epochs* = 50, *layers* = 4, *hidden size* = 128, *batch size* = 128, *learning rate* = 1e-4, *dropout* = 0.5.

PLMs and SVM. We use consistent hyperparameters across tasks, with minor adjustments for PLMs detailed in Table 4. For SVM, we use a linear kernel with a maximum of 60 iterations and a tolerance of 0.01 on both datasets.

LLM Setup. Due to input length constraints, articles are first summarized using BART (Lewis et al., 2020) to 250-300 words. For each outlet, we select five articles that focus on political, economic, and social topics. These summaries are then used as inputs to LLMs through task-specific prompts. We sample 100 outlets from ACL-2020 and 300 from MBFC-2025 for zero- and few-shot (1, 3, 5) evaluation. For few-shot settings, examples are selected to ensure label diversity.

The final outlet-level predictions are obtained using both soft and hard voting on five article-level predictions. Summarization provides additional context, while voting reduces variance and bias, resulting in more stable media-level classification.

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

The summarized article (replacing **{article}**) is then used as evidence for downstream classification.

```
system_prompt = '''You are an expert in media
↪  analysis. Classify the factual reporting
↪  level of the given news article from {media}
↪  into one of the following:
- high
- mixed
- low
Return -1 if uncertain.'''
user_prompt = "{article}"
```

```
system_prompt = '''You are an expert in media
↪  analysis. Classify the political bias of the
↪  given news article from {media} into one of
↪  the following:
- left
- center
- right
Return -1 if uncertain.'''
user_prompt = "{article}"
```

We replace **{media}** placeholder with the outlet name, and **{article}** with the summarized article. For MBFC-2025, we extend the label space to include finer-grained categories: *very high* and *very low* for factuality, and *left-center* and *right-center* for political bias.

RL Agent. We implemented the RL agent using PPO. To model the problem as a contextual bandit, we set the discount factor to 0, ensuring optimization focuses on immediate rewards. The policy network is an MLP with two hidden layers of 128 neurons and `tanh` activations. We use a learning rate of 1×10^{-4}, a batch size of 256, and a rollout size of 1024 environment steps per policy update.

Evaluation Measures. We use MAE, Macro-F1, Accuracy, Precision, and Recall, with MAE reflecting the ordinal nature of labels (Baly et al., 2018, 2020a). Hyperparameters for GNNs and PLMs are tuned on development sets using common values to balance performance and efficiency. All experiments run on an NVIDIA RTX A6000 48GB GPU.

Models	Representation	MAE	Macro-F1	Accuracy	Precision	Recall
ACL-2020						
GREENER (Panayotov et al., 2022)	$Twitter + YouTube + Alexa + A + W$	-	91.93	92.08	-	-
MGM (Manzoor et al., 2025)	$A + Graph\ Embeddings$	-	93.08	93.45	-	93.19
Majority class		0.633	17.91	36.73	12.24	33.33
$\text{RoBERTa}_{\text{Base}}$ Soft Voting	A	0.285	76.16	77.55	80.51	75.91
$\text{RoBERTa}_{\text{Base}}$	W	0.484	66.33	70.47	66.81	66.38
$\text{RoBERTa}_{\text{Base}}$ Soft Voting	$A + W$	0.224	81.47	81.63	83.61	81.20
SVM	$Alexa_{ResGGCN}$	0.596	60.54	61.70	60.67	60.78
SVM	$Alexa_{ResGGCN} + A$	0.106	91.56	91.49	91.61	91.61
Co-attention	$Alexa_{GCN} + Hyperlink_{SAGE} + A$	0.111	93.20	93.33	94.12	93.28
SVM	$Alexa_{ResGGCN} + Hyperlink_{SAGE} + A + W$	**0.067**	95.53	95.56	95.83	95.66
SVM	$LLM_{GCN} + Alexa_{ResGGCN} + Hyperlink_{SAGE} + A + W$	0.089	93.46	93.33	94.12	93.70
RL (PPO)	$Alexa_{GCN} + Hyperlink_{SAGE} + LLM_{GCN} + A + W$	0.098	**97.60**	**97.60**	**97.30**	**98.00**
MBFC-2025						
Majority class		1.071	9.21	29.92	5.98	20.00
$\text{ALBERT}_{\text{Base}}$ Soft Voting	A	0.457	71.74	74.41	71.84	71.61
Ensemble Hard Voting	W	0.623	63.19	64.15	65.33	62.86
Ensemble Soft Voting	$A + W$	0.425	72.41	75.47	73.69	72.36
SVM	$Alexa_{GCN}$	1.010	40.44	46.08	42.59	41.94
SVM	$A + W$	0.405	69.09	73.14	71.00	68.67
SVM	$Hyperlink_{GCN} + A + W$	**0.395**	69.29	73.18	70.81	69.05
MLP	$LLM_{GCN} + Hyperlink_{SAGE} + A + W$	0.482	67.04	70.91	66.93	68.46
Self-attention	$LLM_{SAGE} + LLM_{GCN} + Hyperlink_{SAGE} + A + W$	0.486	67.79	70.00	68.53	71.09
RL (PPO)	$Alexa_{GCN} + Hyperlink_{SAGE} + LLM_{GCN} + A + W$	0.551	**78.00**	**77.90**	**77.80**	**78.30**

Table 5: Evaluation results using multiple representations for *political bias* detection and baselines from prior work. The representation column indicates the embeddings used as input: Articles (A), Wikipedia (W), Alexa graph ($Alexa$), Hyperlink graph ($Hyperlink$), and LLM-based graph (LLM). Subscripts denote the GNN encoder. **Green** marks the best scores, while **Blue** marks the lowest scores.

5.2 Evaluation

We discuss our best results in § 5.2.1, with additional results in Appendix C. The model analysis is provided in § 5.2.3, and key findings are summarized in § 5.2.4.

5.2.1 Results

As illustrated in Table 5 for *political bias*, the combination of media Articles and Wikipedia representations, together with Alexa graph (GCN) and Hyperlink graph (GraphSAGE), using an SVM achieved the strongest results among all static baselines on ACL-2020, with a Macro-F1 of 95.53%.

Moreover, our RL-based fusion strategy consistently outperformed all static approaches and achieved a higher Macro-F1 of 97.60%. On the larger and more challenging MBFC-2025, our RL-based dynamic fusion strategy again significantly outperformed all static baselines and ensemble methods. Although the best static fusion achieved a Macro-F1 of 69.29% and ensemble voting reached 72.41%, our RL agent achieved a new state-of-the-art Macro-F1 of **78.00%**. These results demonstrate the efficacy of dynamic weight assignment, particularly in complex, multi-view settings where static fusion strategies performed poorly.

As shown in Table 7 for *factuality*, we combine the Hyperlink graph (GCN) and the articles.

Textual embeddings fused via cross-attention emerged as the strongest baseline, achieving a Macro-F1 of 78.63% and 84.44% accuracy. Our RL-based fusion further raised accuracy to 86.60% while maintaining a comparable Macro-F1 of 76.80%, indicating reliable instance-level predictions without sacrificing class balance. On MBFC-2025, the RL agent achieved 55.30% Macro-F1, performing on par with other fusion techniques such as Co-attention (56.94%). However, text-only models remained dominant for the factuality task overall, with $\text{ALBERT}_{\text{Base}}$ reaching 71.74% Macro-F1 through hard voting.

5.2.2 GNN Embeddings for SVM

For *factuality* (Table 7), SVM with the Alexa graph in ACL-2020 achieved the best Macro-F1 of 42.06% using GCN, while ResGatedGCN yielded the lowest MAE of 0.532. In MBFC-2025, LLM-based embeddings with GraphSAGE achieved the lowest MAE (0.365), while the Alexa graph with GCN gave the best Macro-F1 (30.76%). For *political bias* (Table 8), SVM with Alexa graph and ResGatedGCN in ACL-2020 achieved a Macro-F1 of 60.54% and an MAE of 0.596. In MBFC-2025, the Alexa graph with GCN had the highest Macro-F1 (40.44%), while LLM-based embeddings with GCN achieved the lowest MAE (0.787).

Models	Representation	MAE	Macro-F1	Accuracy	Precision	Recall
ACL-2020						
Node classification (NC) (Mehta et al., 2022)	$Twitter + YouTube + SocialGraph + UsersProfiles + A$	-	68.90	63.72	-	-
InfOp Best Model (Mehta et al., 2022)	$Twitter + YouTube + SocialGraph + UsersProfiles + A$	-	72.55	66.89	-	-
GREENER (Panayotov et al., 2022)	$Twitter + YouTube + Alexa + A + W$	-	69.61	74.27	-	-
MGM (Manzoor et al., 2025)	$A + Graph\ Embeddings$	-	**79.72**	84.21	-	76.54
Majority class		0.571	24.79	59.18	19.73	33.33
$ALBERT_{Base}$ Hard Voting	A	0.449	51.46	67.35	76.14	50.72
$SVM_{TF\text{-}IDF}$	W	0.551	36.92	57.14	34.12	40.33
$BERT_{Base}$	W	0.531	30.20	61.22	37.23	36.11
$ALBERT_{Base}$ Hard Voting	$A + W$	0.469	42.42	65.31	77.04	43.06
SVM	$Alexa_{GCN}$	0.574	42.06	55.32	49.30	42.49
Cross-attention	$Hyperlink_{GCN} + A$	**0.178**	78.63	84.44	81.25	**77.59**
MLP	$Alexa_{ResGGCN} + Hyperlink_{SAGE} + A$	**0.178**	76.07	82.22	78.86	74.82
Self-attention	$LLM_{GCN} + Hyperlink_{SAGE} + Hyperlink_{GCN} + A$	0.267	73.40	77.78	74.60	74.73
Cross-attention	$LLM_{GCN} + Alexa_{GCN} + Hyperlink_{SAGE} + Hyperlink_{GCN} + A$	0.222	74.63	80.00	75.79	76.01
RL (PPO)	$Alexa_{GCN} + Hyperlink_{SAGE} + LLM_{GCN} + A + W$	0.183	76.80	**86.60**	**91.20**	71.70
MBFC-2025						
Majority class		0.567	13.81	52.76	10.55	20.00
$ALBERT_{Base}$ Hard Voting	A	0.457	**71.74**	**74.41**	**71.84**	**71.61**
$RoBERTa_{Base}$	W	0.413	29.49	63.68	45.47	30.04
$DistilBERT_{Base}$	W	0.398	27.45	64.68	26.07	29.18
$RoBERTa_{Base}$ Hard Voting	$A + W$	0.313	41.58	71.64	68.58	38.63
SVM	$Alexa_{GCN}$	0.405	30.76	65.29	44.33	29.94
Cross-attention	$A + W$	**0.277**	51.78	75.62	49.33	58.98
Co-attention	$Alexa_{ResGGCN} + A + W$	0.291	56.94	74.36	55.06	61.74
Cross-attention	$LLM_{GCN} + Hyperlink_{SAGE} + A + W$	0.306	50.82	71.76	51.35	61.72
Cross-attention	$LLM_{SAGE} + LLM_{GCN} + Hyperlink_{GCN} + A + W$	0.310	53.30	71.76	50.90	64.32
RL (PPO)	$Alexa_{GCN} + Hyperlink_{SAGE} + LLM_{GCN} + A + W$	0.249	55.30	72.10	67.10	62.80

Table 6: Evaluation results using multiple representations for *factuality* detection and baselines from prior work. Representations include Articles (A), Wikipedia (W), Alexa graph ($Alexa$), Hyperlink graph ($Hyperlink$), and LLM-based graph (LLM). Subscripts denote the GNN encoder. **Green** marks the best scores, while **Blue** marks the lowest scores.

5.2.3 Discussion

As shown in Tables 5 and 6, we evaluated combinations of representations to detect *factuality* and *political bias* in various methods and datasets.

GNNs. As detailed in Section 5.2.2, there is no particular trend in GNN performance. For ACL-2020, SVM+Alexa rank features yield the best factuality prediction results; for MBFC-2025, LLM+GraphSAGE yields the lowest MAE, and Alexa+GCN yields the best Macro-F1. For political bias in the ACL dataset, the best combination was SVM+Alexa + ResGatedGCN, while for MBFC-2025, Alexa+GCN gave the best Macro-F1, and LLM+GraphSAGE gave the lowest MAE 0.203.

PLMs and LLMs. PLMs and LLMs generally predict political bias better than factuality, with Wikipedia pages having a stronger influence on LLMs. Fine-tuned PLMs on both media articles and Wikipedia descriptions outperform those trained on Wikipedia alone. However, combining representations in LLMs reduces performance compared to a single representation.

5.2.4 Empirical Findings

(*i*) Political bias is easier to detect than factuality, as it is more explicit in language, while factuality requires deeper contextual understanding and implicit verification (Panayotov et al., 2022).

(*ii*) Single-view graph representations are suboptimal and combining multiple views yields stronger results by capturing diverse network dependencies. (*iii*) PLMs benefit from multiple representations, whereas LLMs often perform best with a single high-quality view, likely because additional views introduce noise and increase optimization complexity. (*iv*) Multi-view evaluation improves predictive performance and interpretability by clarifying which representations and interactions matter most. However, as shown in Appendix D, adding more representations does not always improve the results, often leading to diminishing returns. (*v*) The Alexa graph is the most effective view due to its broad coverage and informative node features. Hyperlink graphs provide complementary signals, whereas the LLM graph is weakest due to missing explicit node features. (*vi*) Performance is higher on ACL-2020 than on MBFC-2025, likely because the 5-point scale of MBFC introduces finer distinctions and greater ambiguity. (*vii*) RL-based fusion improves political bias detection, but not strong factuality prediction, as political bias signals are more consistent between views and support more reliable reward-driven selection. In contrast, factuality is more instance-dependent and requires stronger grounding, leading to noisier rewards and a policy more susceptible to spurious correlations.

Representation	MAE	Macro-F1	Accuracy	Precision	Recall
	ACL-2020				
LLM_{SAGE}	0.596	27.64	55.32	35.19	34.43
LLM_{GCN}	0.596	24.07	55.32	18.84	33.33
$Alexa_{ResGGCN}$	**0.532**	40.53	**57.45**	46.12	41.39
$Alexa_{GCN}$	0.574	**42.06**	55.32	**49.30**	**42.49**
$Hyperlink_{SAGE}$	0.689	37.70	51.11	41.25	37.45
$Hyperlink_{GCN}$	0.600	28.33	55.56	30.62	33.55
	MBFC-2025				
LLM_{SAGE}	**0.365**	27.99	**68.25**	27.00	29.28
LLM_{GCN}	0.373	28.03	67.46	26.80	29.39
$Alexa_{ResGGCN}$	0.438	26.42	62.81	27.79	27.34
$Alexa_{GCN}$	0.405	**30.76**	65.29	**44.33**	**29.94**
$Hyperlink_{SAGE}$	0.489	23.33	59.39	23.21	23.93
$Hyperlink_{GCN}$	0.476	22.12	61.57	23.65	23.60

Table 7: Evaluation results using SVM with different GNN embeddings for *factuality* detection across ACL-2020 and MBFC-2025 datasets.

Representation	MAE	Macro-F1	Accuracy	Precision	Recall
	ACL-2020				
LLM_{SAGE}	0.872	31.03	36.17	27.12	36.69
LLM_{GCN}	0.809	35.14	40.43	59.87	40.74
$Alexa_{ResGGCN}$	**0.596**	**60.54**	**61.70**	**60.67**	**60.78**
$Alexa_{GCN}$	**0.596**	57.50	57.45	57.64	57.70
$Hyperlink_{SAGE}$	0.911	23.92	24.44	23.93	24.09
$Hyperlink_{GCN}$	0.667	41.49	42.22	45.42	41.46
	MBFC-2025				
LLM_{SAGE}	0.802	26.77	43.48	28.97	31.56
LLM_{GCN}	**0.787**	25.39	44.27	25.94	31.48
$Alexa_{ResGGCN}$	1.088	38.61	43.14	38.45	39.54
$Alexa_{GCN}$	1.010	**40.44**	**46.08**	**42.59**	**41.94**
$Hyperlink_{SAGE}$	0.957	31.15	40.69	36.35	32.57
$Hyperlink_{GCN}$	0.870	27.62	39.83	38.17	30.36

Table 8: Evaluation results using SVM with different GNN embeddings for *political bias* detection across ACL-2020 and MBFC-2025 datasets.

6 Conclusion and Future Work

We introduced **MBFC-2025**, a large-scale, fine-grained annotation set, and constructed multi-view representations across two benchmark datasets. Through an extensive evaluation of embedding combinations and fusion strategies, including an RL-based dynamic fusion variant, we achieve state-of-the-art results in ACL-2020 and establish strong benchmarks in MBFC-2025, where the RL agent improves over static fusion methods, particularly for political-bias detection. We aim to expand our dataset into a larger multilingual corpus that captures political bias across diverse cultures and regions, thereby reducing the current U.S.-centric focus.

In future work, we also plan to scale our modeling approach by incorporating graph neural networks (GNNs) and parameter-efficient fine-tuning of large language models, using methods such as LoRA (Hu et al., 2022), LoFT (Tastan et al., 2026b), and potentially mixture-of-expert architectures (Tastan et al., 2026a).

Limitations

This study has data coverage limitations. Despite efforts to collect data from MBFC-annotated outlets, some websites could not be parsed. We also used Wikipedia for descriptive text and did not compare its labels with MBFC political leanings due to inconsistent availability and difficulty at scale. A further limitation concerns graph construction. The hyperlink and LLM graphs lack rich node attributes, so GNNs were trained using dummy features (e.g., a single integer per node). This likely constrained performance, although the embeddings still captured useful structural patterns.

Also, our experiments are limited by computational and modeling constraints. Due to time and memory requirements, we prioritized efficient models and did not explore fine-tuning, leaving it for future work. These constraints also prevented evaluation of larger LLMs.

More broadly, our framework relies mainly on U.S.-centric political categories such as left, center, and right, which may not fully capture ideological variation in other cultural or political settings. The dataset should therefore not be used for fine-grained political bias or factuality detection, or for article-level analysis without broader context. Our work focuses on source-level political bias as a starting point, and future research should extend this setting to article- and claim-level prediction. Finally, while this study focuses on classification, it does not yet examine broader structural patterns in media ecosystems or more advanced fusion strategies, both of which remain important directions for future work.

Ethical Statement & Bias

Articles were collected according to legal and ethical standards, using publicly available content, adhering to site policies, and limiting access frequency to reduce server load. No paid or restricted content was accessed. Despite these safeguards, the data may still contain biases: Wikipedia-derived information can reflect implicit leanings, and both media sources and annotations may introduce systematic bias. Although, we include a diverse range of outlets to mitigate these effects, the data may still underrepresent certain perspectives. As a result, findings should be interpreted with caution, particularly when generalizing beyond the sources studied or applying models in real-world settings.

Moreover, automated profiling and summarization may inadvertently reinforce polarization if users rely on condensed output rather than full context. Consequently, the framework is designed for transparency and research analysis, not for content filtering or personalization.

To protect anonymity and data integrity, news articles and graph representations are not publicly available; instead, only the code, scraping recipes, and relevant URLs (including article/media links and labels) are provided to support reproducibility.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sadia Afroz, Michael Brennan, and Rachel Greenstadt. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proc. of the 2012 IEEE Symposium on Security and Privacy*, page 461–475, San Francisco, CA, USA. IEEE Computer Society.

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 Assoc. for Comp. Linguistics: EMNLP 2024*, pages 12217–12231, Miami, Florida, United States.

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*, CEUR Workshop Proceedings, pages 289–305. CEUR-WS.org.

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'20, pages 4982–4991, Online. Association for Computational Linguistics.

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*, EMNLP'18, 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. 2020b. 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 Assoc. for Comp. Linguistics*, ACL'20, 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 Assoc. for Comp. Linguistics: Human Language Technologies*, NAACL'19, 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, Gullal S. Cheema, Dilshod Azizov, and Preslav Nakov. 2023a. The CLEF-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III*, Lecture Notes in Computer Science, 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, Maram Hasanain, Mücahid Kutlu, Chengkai Li, Federico Ruggeri, Julia Maria Struß, and Wajdi Zaghouani. 2023b. 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 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, Lecture Notes in Computer Science, pages 251–275. Springer.

Cristian Bodnar, Fabrizio Frasca, Yuguang Wang, Nina Otter, Guido F. Montúfar, Pietro Lió, and Michael M. Bronstein. 2021. Weisfeiler and lehman go topological: Message passing simplicial networks. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research, pages 1026–1037. PMLR.

Michael P Boyle, Mike Schmierbach, and Douglas M McLeod. 2007. Ideology, issues, and limited information: Implications for voting behavior. *Atlantic Journal of Communication*, 15(4):284–302.

Xavier Bresson and Thomas Laurent. 2017. Residual gated graph ConvNets. *arXiv preprint arXiv:1711.07553*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33 of *NeurIPS'20*, pages 1877–1901, Vancouver, BC, Canada. Curran Associates, Inc.

Sergio Burdisso, Dairazalia Sanchez-cortes, Esaú Villatoro-tello, and Petr Motlicek. 2024. Reliability estimation of news media sources: Birds of a feather flock together. In *Proceedings of the 2024 Conference of the North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL'24, pages 6893–6911, Mexico City, Mexico. Association for Computational Linguistics.

Sonia Castelo, Thais G. Almeida, Anas Elghafari, Aécio S. R. Santos, Kien Pham, Eduardo Freire Nakamura, and Juliana Freire. 2019. A topic-agnostic approach for identifying fake news pages. In *Companion of The 2019 World Wide Web Conference*, WWW'19, pages 975–980, San Francisco, CA, USA. ACM.

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.

Ganqu Cui, Jie Zhou, Cheng Yang, and Zhiyuan Liu. 2020. Adaptive graph encoder for attributed graph embedding. In *Proceedings of the 26th International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 976–985, Virtual Event, CA, USA. Association for Computing Machinery.

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.

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 Assoc. for Comp. Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL'19, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaize Ding, Zhe Xu, Hanghang Tong, and Huan Liu. 2022. Data augmentation for deep graph learning: A survey. *SIGKDD Explorations Newsletter*, 24(2):61–77.

James N Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049.

James Fairbanks, Natalie Fitch, Nathan Knauf, and Erica Briscoe. 2018. Credibility assessment in the news: do we need to read? In *Proceedings of the MIS2 Workshop on 11th Int. Conf. on Web Search and Data Mining*, WSDM'18, pages 799–800. ACM.

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.

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*, ICWSM'22, pages 1404–1408.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 30 of *NeurIPS'17*, page 1025–1035, Long Beach, CA, USA. Curran Associates Inc.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying disinformation websites using infrastructure features. In *Proceedings of the 10th USENIX Workshop on Free and Open Communications on the Internet*, FOCI'20, Virtual.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, ICLR '22, virtual.

Noussaiba Jaafar and Zied Lachiri. 2023. Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, 211:118523.

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*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations*, ICLR'20.

Yuanyuan Lei and Ruihong Huang. 2024. Sentence-level media bias analysis with event relation graph. In *Proceedings of the 2024 Conference of the North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies (Volume 1:*

Long Papers), NAACL'24, pages 5225–5238, Mexico City, Mexico. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Assoc. for Comp. Linguistics*, ACL'20, pages 7871–7880, Online. Association for Computational Linguistics.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024. IndiVec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Findings of the Assoc. for Comp. Linguistics: EACL 2024*, pages 1038–1050, St. Julian's, Malta. 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*, abs/1907.11692.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Assoc. for Comp. Linguistics: NAACL 2022*, Seattle, United States. Association for Computational Linguistics.

Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Liò, Bruno Lepri, and Andrea Passerini. 2024. Explaining the explainers in graph neural networks: a comparative study. *ACM Computing Surveys*.

Hehuan Ma, Yatao Bian, Yu Rong, Wenbing Huang, Tingyang Xu, Weiyang Xie, Geyan Ye, and Junzhou Huang. 2020. Multi-view graph neural networks for molecular property prediction. *arXiv preprint arXiv:2005.13607*.

Iffat Maab, Edison Marrese-Taylor, Sebastian Padó, and Yutaka Matsuo. 2024. Media bias detection across families of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL'24, pages 4083–4098, Mexico City, Mexico. Association for Computational Linguistics.

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 Assoc. for Comp. Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7279–7295, Albuquerque, New Mexico, USA.

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

Nikhil Mehta, Maria 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 Assoc. for Comp. Linguistics (Volume 1: Long Papers)*, ACL'22, pages 1363–1380, Dublin, Ireland. Association for Computational Linguistics.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and Leman go neural: higher-order graph neural networks. In *Proceedings of the Thirty-Third Conference on Artificial Intelligence*, AAAI'19, Honolulu, Hawaii, USA. AAAI Press.

Zain Muhammad Mujahid, Dilshod Azizov, Maha Tufail Agro, and Preslav Nakov. 2025. Profiling news media for factuality and bias using LLMs and the fact-checking methodology of human experts. In *Findings of the Assoc. for Comp. Linguistics: ACL 2025*, pages 798–819, Vienna, Austria. ACL.

Subhabrata Mukherjee and Gerhard Weikum. 2015. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th International Conference on Information and Knowledge Management*, CIKM'15, pages 353–362, Melbourne, VIC, Australia. ACM.

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 Assoc. for Comp. Linguistics: ACL 2024*, pages 15947–15962, Bangkok, Thailand. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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*, volume 35 of *NeurIPS'22*, pages 27730–27744, New Orleans, Louisiana, USA. Curran Associates, Inc.

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, pages 7470–7480, Abu Dhabi, United Arab Emirates. ACL.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, CCL'18, pages 3391–3401, Santa Fe, New Mexico, USA. ACL.

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, Perth, Australia.

Markus Prior. 2013. Media and political polarization. *Annual review of political science*, 16(1):101–127.

Changyuan Qiu, Winston Wu, Xinliang Frederick Zhang, and Lu Wang. 2022. Late fusion with triplet margin objective for multimodal ideology prediction and analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, EMNLP'22, pages 9720–9736, Abu Dhabi, United Arab Emirates. ACL.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.

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*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Ke Sun, Zhouchen Lin, and Zhanxing Zhu. 2020. Multi-stage self-supervised learning for graph convolutional networks on graphs with few labeled nodes. In *Proceedings of the Conference on Artificial Intelligence*, volume 34 of *AAAI'20*, pages 5892–5899.

Nurbek Tastan, Stefanos Laskaridis, Karthik Nandakumar, and Samuel Horvath. 2026a. MoSE: Mixture of slimmable experts for efficient and adaptive language models. *arXiv preprint arXiv:2602.06154*.

Nurbek Tastan, Stefanos Laskaridis, Martin Takáč, Karthik Nandakumar, and Samuel Horváth. 2026b. LoFT: Low-rank adaptation that behaves like full fine-tuning. In *The Fourteenth Int. Conf. on Learning Representations*, ICLR'26, Rio de Janeiro, Brazil.

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*.

Mark Towers, Ariel Kwiatkowski, Jordan K. Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. 2024. Gymnasium: A standard interface for reinforcement learning environments. *CoRR*, abs/2407.17032.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22, New Orleans, LA, USA. Curran Associates Inc.

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*, Websci '25, page 127–137, New Brunswick, New Jersey, USA. ACM.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23, New Orleans, LA, USA. Curran Associates Inc.

Jinliang Yuan, Hualei Yu, Meng Cao, Ming Xu, Junyuan Xie, and Chongjun Wang. 2021. Semi-supervised and self-supervised classification with multi-view graph neural networks. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2466–2476, Virtual Event. ACM.

Jiahao Zhang, Rui Xue, Wenqi Fan, Xin Xu, Qing Li, Jian Pei, and Xiaorui Liu. 2024. Linear-time graph neural networks for scalable recommendations. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3533–3544, Singapore, Singapore. Association for Computing Machinery.

Wenqian Zhang, Shangbin Feng, Zilong Chen, Zhenyu Lei, Jundong Li, and Minnan Luo. 2022. KCD: Knowledge walks and textual cues enhanced political perspective detection in news media. In *Proceedings of the 2022 Conference of the North American Chapter of the Assoc. for Comp. Linguistics: Human Language Technologies*, NAACL'22, pages 4129–4140, Seattle, United States. ACl.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, and Quoc V. Le. 2023. Least-to-most prompting enables complex reasoning in large language models. In *Proceedings of the International Conference on Learning Representations*, ICLR '23, Kigali, Rwanda.

A Data Statement for MBFC-2025

Dataset Version 1.0 (October 2024)

Data Statement Version 1.0 (October 2024)

Data Collection Period The dataset was collected from February to October 2024.

A.1 Executive Summary: MBFC-2025 is a dataset focused on analyzing political bias and factuality in English-language media at the outlet level. The dataset includes *Articles* from various news media outlets, combined with *Wikipedia* descriptions and graph-based representations. This dataset is designed to support research in media analysis.

A.2 Granularity: The dataset analysis is presented on a 5-point scale for political bias (Left, Left-Center, Center, Right-Center, Right) and factuality (Very High, High, Mixed, Low, Very Low).

A.3 Documentation for Source Datasets: The dataset was sourced from a variety of English-language media outlets, annotated by Media Bias/Fact Check experts. Data were collected through a systematic process that involved extraction of *Articles* from news media websites and their corresponding *Wikipedia* descriptions. Graph-based representations were obtained using the Alexa graph tool, Hyperlink graph, and LLM. The selection of sources was guided by relevance, credibility, and the objective of covering a broad spectrum of political bias and factuality.

A.4 Annotation Process: Political bias and factuality annotations were conducted by experts from Media Bias/Fact Check following a standardized protocol.

A.5 Intended Use: The dataset is intended for research on political bias and factuality at the media-level. It is particularly suited for tasks such as the detection of political bias and factuality, media analysis, and related computational studies.

A.7 Data Collection: Label Collection from MBFC: The labels were scraped from MBFC's website using a web crawler, extracting political bias and factuality annotations for each media. To ensure accuracy, we parsed the HTML structure, identified classification tags, structured the data, and manually verified labels and media alignments to confirm correct extraction.

Articles: To collect articles, we focus on sections that cover political, economic, and social issues, selecting a variety of specific topics within these areas. The process involved two main steps: *(i)* we exclusively used media sources annotated by experts from Media Bias/Fact Check (MBFC), and *(ii)* we selected active media outlets. From these sources, we collected up to 30 front-page articles from each website.

The data collection process consisted of four stages: *(i)* we gathered media sources from MBFC, extracting each source's details as JSON lines from the HTML code after manually verifying accessibility through its links, *(ii)* we parsed front-page article links, excluding menu links and collecting only internal domain links over 65 characters, *(iii)* the titles and article texts were retrieved using a combination of script automation and manual testing to ensure successful extraction, and *(iv)* the post-processing stage involved formatting the data into the required JSON format.

Wikipedia: To locate the *Wikipedia* link, we first searched for the name of the media outlet online. We ensure that the link points to a *Wikipedia* entry about the media outlet. Next, we extracted the webpage in its uniform HTML format from the *Wikipedia* website. Then we parsed the HTML to extract the page content. The last step of the post-processing phase was to transform the gathered data into the necessary JSON format.

B Statistics of Alexa Media Graph

Table 9 presents the statistics for the media graph constructed with the Alexa Rank tool based on audience overlap. The graph levels correspond to iterative expansions conducted by the authors (Panayotov et al., 2022) to enrich the graph and improve the contextual representation of the media within the network. The table clearly shows that as we progress to higher levels, both the number of nodes and edges increase, while the number of graph components decreases, which is beneficial for GNNs in learning effective node representations. However, at the highest level (Level 3), there are still 44 disconnected components, which may hinder the GNNs' ability to learn or generalize, potentially leading to sub-optimal node representations and lower performance on classification tasks.

Level	Nodes	Edges	# of components	Avg. # of nodes
0	4563	20210	326	10.7
1	10161	28779	142	71.55
2	26573	78600	75	354.3
3	67351	200488	44	1530.7

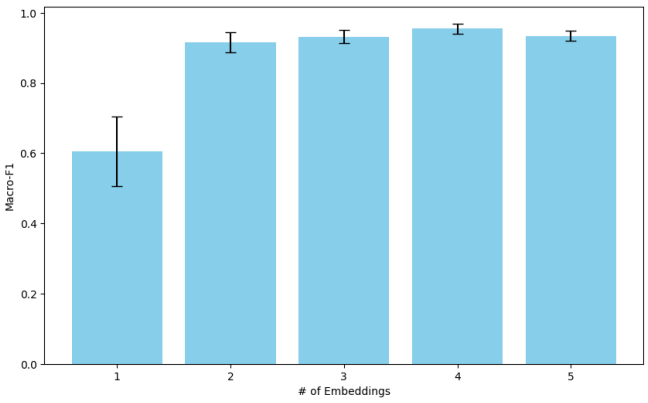
Table 9: Graph statistics across different levels.

C Results

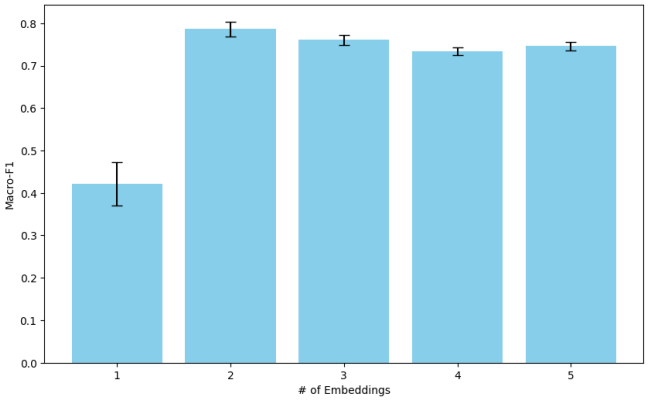
Our experimental results are shown in Tables 10-16. Table 10 shows the dummy-classifier baseline results for *political bias* and *factuality* on ACL-2020 and MBFC-2025. Table 11 shows the GPT-4o hard-voting results across the same two tasks and datasets. Table 12 shows the article-level *political bias* classification results using hard and soft voting for PLMs, LLMs, and their ensembles. Table 13 shows the corresponding article-level *factuality* classification results. Table 14 shows the media-level results on Wikipedia-based datasets for both *political bias* and *factuality*, including independent frameworks and ensemble models under hard-voting and soft-voting settings. Finally, Table 15 and Table 16 show the *political bias* and *factuality* results, respectively, for the third experimental setting, again comparing PLMs, LLMs, and ensemble methods under both voting strategies.

D Dependency on Representations

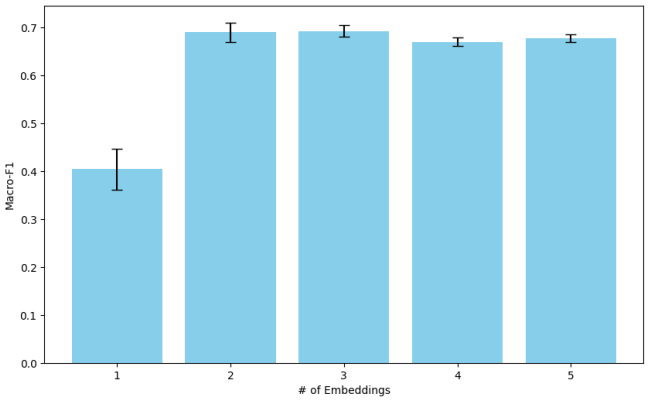
In Figure 4 we illustrate how the number of representations affects the results. Overall, 2-3 representations are mainly enough; adding more often results in considerably lower gains.



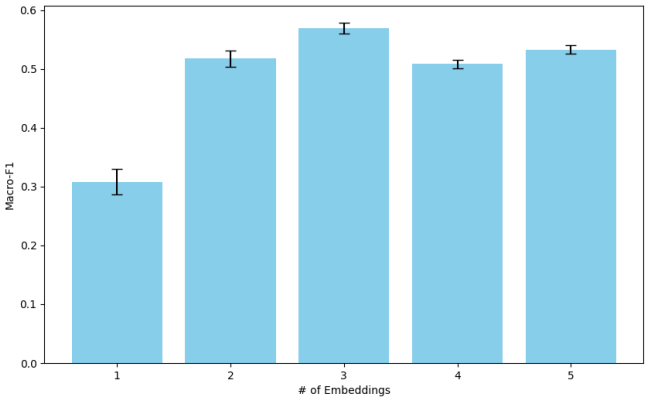
(a) Political Bias - ACL2020.



(b) Factuality - ACL2020.



(c) Political Bias - MBFC2024.



(d) Factuality - MBFC2024.

Figure 4: Dependency of *Macro-F1* on the number of representations, with a 0.9 confidence interval.

Models	Political Bias					Factuality				
	MAE	**Macro-F1**	**Accuracy**	**Precision**	**Recall**	**MAE**	**Macro-F1**	**Accuracy**	**Precision**	**Recall**
	ACL-2020									
Majority class	**0.633**	**17.91**	**36.73**	**12.24**	**33.33**	**0.571**	**24.79**	**59.18**	**19.73**	**33.33**
Middle class	1.061	14.81	28.57	9.52	**33.33**	1.429	9.36	16.33	5.44	**33.33**
	MBFC-2025									
Majority class	**1.071**	**9.21**	**29.92**	**5.98**	**20.00**	**0.567**	**13.81**	**52.76**	**10.55**	**20.00**
Middle class	1.449	5.68	16.54	3.31	**20.00**	1.504	2.09	5.51	1.10	**20.00**

Table 10: Evaluation results for *political bias* and *factuality* using dummy classifiers for the majority and middle classes in the ACL-2020 and MBFC-2025 datasets.

Dataset	Model	MAE	Macro-F1	Accuracy	Precision	Recall
ACL-2020 Bias	GPT-4o$_0$	0.562	62.50	54.37	60.83	49.33
	GPT-4o$_1$	0.250	77.08	59.88	60.00	60.49
	GPT-4o$_3$	**0.229**	**79.17**	**61.40**	**60.86**	**62.28**
	GPT-4o$_5$	0.250	77.08	59.88	60.00	60.49
ACL-2020 Factuality	GPT-4o$_0$	0.625	39.58	36.77	60.83	45.41
	GPT-4o$_1$	**0.766**	**42.86**	**35.69**	**33.02**	**43.43**
	GPT-4o$_3$	0.667	39.58	38.19	55.23	46.65
	GPT-4o$_5$	0.729	29.17	24.37	47.44	32.28
MBFC-2025 Bias	GPT-4o$_0$	**0.766**	**42.86**	**35.69**	**33.02**	**43.43**
	GPT-4o$_1$	2.065	11.69	4.19	2.34	20.00
	GPT-4o$_3$	2.065	11.69	4.19	2.34	20.00
	GPT-4o$_5$	2.065	11.69	4.19	2.34	20.00
MBFC-2025 Factuality	GPT-4o$_0$	1.468	9.09	6.32	10.62	16.78
	GPT-4o$_1$	**1.416**	**10.39**	**6.67**	8.20	17.47
	GPT-4o$_3$	2.169	3.90	4.62	4.66	**27.36**
	GPT-4o$_5$	2.494	2.60	1.85	**20.26**	20.69

Table 11: Results for *political bias* and *factuality* on ACL-2020 and MBFC-2025 media articles datasets using GPT-4o hard voting. The subscript denotes the number of few-shot examples used in the prompt.

Models	Hard Voting					Soft Voting				
	MAE	**Macro-F1**	**Accuracy**	**Precision**	**Recall**	**MAE**	**Macro-F1**	**Accuracy**	**Precision**	**Recall**
					ACL-2020					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.571	37.00	44.89	63.49	42.54	0.551	33.86	44.89	46.66	42.85
$BERT_{Base}$	0.346	69.03	69.38	72.44	68.39	0.367	68.18	69.38	72.62	67.86
$RoBERTa_{Base}$	0.306	75.93	**77.55**	75.91	**75.93**	**0.285**	**76.16**	**77.55**	**80.51**	75.91
$DistilBERT_{Base}$	0.346	68.54	69.38	72.59	67.97	0.306	72.99	73.46	77.62	72.20
$ALBERT_{Base}$	0.510	62.12	65.30	64.39	63.10	0.428	67.05	69.38	70.41	67.44
Ensemble	0.416	63.52	66.66	71.82	64.21	0.416	63.52	66.66	71.82	64.21
					LLMs					
LLaMA2 $_0$	1.933	11.05	11.94	11.15	13.13	1.930	11.59	24.17	16.44	24.17
LLaMA2 $_1$	1.314	28.39	38.74	22.83	38.74	1.307	29.10	39.74	23.53	39.74
LLaMA2 $_3$	**1.201**	34.42	41.72	34.88	41.72	1.182	**36.41**	**44.37**	35.85	**44.37**
LLaMA2 $_5$	1.314	24.13	32.45	35.32	32.45	1.317	24.57	33.11	37.04	33.11
Mistral $_0$	1.923	9.53	23.51	7.47	23.51	1.894	12.49	25.17	23.86	25.17
Mistral $_1$	1.390	12.34	21.19	35.14	21.19	1.384	14.45	22.52	37.12	22.52
Mistral $_3$	1.357	18.10	25.17	35.13	25.17	1.341	20.79	27.15	**36.52**	27.15
Mistral $_5$	1.417	7.01	18.21	22.87	18.21	1.417	7.01	18.21	22.87	18.21
Ensemble $_0$	1.930	9.58	23.51	15.47	23.51	1.943	9.44	23.51	15.36	23.51
Ensemble $_1$	1.407	22.43	28.48	22.69	28.48	1.205	35.19	44.37	34.46	44.37
Ensemble $_3$	1.278	27.35	32.78	36.28	32.78	1.238	31.79	38.08	37.14	38.08
Ensemble $_5$	1.387	15.78	23.84	33.12	23.84	1.324	21.85	28.81	36.05	28.81
					MBFC-2025					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.630	50.80	58.66	60.65	50.20	0.591	47.74	57.87	65.65	47.28
$BERT_{Base}$	0.539	63.85	68.50	65.55	63.11	0.571	61.51	65.75	62.98	60.66
$RoBERTa_{Base}$	0.484	66.33	70.47	66.81	66.38	0.528	64.51	69.69	65.58	64.53
$DistilBERT_{Base}$	0.547	63.95	68.90	66.74	63.35	0.583	62.27	67.32	64.45	61.59
$ALBERT_{Base}$	0.484	68.93	72.05	69.68	69.00	**0.457**	**71.74**	**74.41**	**71.84**	**71.61**
Ensemble	0.465	69.48	72.44	70.37	69.20	0.472	68.71	72.05	69.48	68.43
					LLMs					
LLaMA2 $_0$	1.933	11.05	23.84	15.33	23.84	1.930	11.59	24.17	**16.44**	24.17
LLaMA2 $_1$	**1.130**	14.38	19.51	14.28	19.51	1.122	14.52	20.33	14.45	20.33
LLaMA2 $_3$	1.789	12.27	21.14	10.61	21.14	1.821	11.87	20.33	11.50	20.33
LLaMA2 $_5$	1.122	**17.17**	22.76	15.84	22.76	1.130	17.13	22.76	15.63	22.76
Mistral $_0$	1.924	9.53	23.51	7.47	23.51	1.894	12.49	25.17	23.86	25.17
Mistral $_1$	1.512	9.71	21.95	10.81	21.95	1.504	10.82	22.76	11.66	22.76
Mistral $_3$	1.480	13.50	23.58	14.35	23.58	1.480	14.61	24.39	16.20	24.39
Mistral $_5$	1.447	11.54	23.58	11.61	23.58	1.439	12.33	24.39	12.27	24.39
Ensemble $_0$	1.930	9.58	23.51	15.47	23.51	1.944	9.44	23.51	15.36	23.51
Ensemble $_1$	1.325	15.41	24.39	15.23	24.39	1.333	15.25	**25.20**	12.62	**25.20**
Ensemble $_3$	1.602	13.27	22.76	9.96	22.76	1.683	14.11	24.39	10.67	24.39
Ensemble $_5$	1.325	12.93	21.95	12.33	21.95	1.211	14.92	21.95	13.76	21.95

Table 12: Evaluation results for *political bias* using hard- and soft-votings for each framework and ensemble in ensemble.

Models	Hard Voting					Soft Voting				
	MAE	**Macro-F1**	**Accuracy**	**Precision**	**Recall**	**MAE**	**Macro-F1**	**Accuracy**	**Precision**	**Recall**
					ACL-2020					
					PLMs					
$SVM_{TF\text{-}IDF}$	**0.449**	48.43	65.31	72.81	47.94	0.551	30.24	61.22	53.47	36.11
$BERT_{Base}$	0.510	32.37	59.18	30.95	36.59	0.551	32.50	59.18	33.79	36.59
$RoBERTa_{Base}$	0.510	34.64	59.18	32.78	38.22	0.510	34.64	59.18	32.78	38.22
$DistilBERT_{Base}$	**0.449**	34.26	63.27	33.59	38.89	0.510	30.22	61.22	32.13	36.11
$ALBERT_{Base}$	**0.449**	**51.46**	**67.35**	**76.14**	**50.72**	0.469	47.16	65.31	74.33	46.31
Ensemble	0.479	34.13	62.50	33.33	38.24	0.479	35.37	64.58	38.64	39.39
					LLMs					
LLaMA2 $_0$	2.095	14.60	23.84	13.74	23.84	2.263	10.17	19.19	12.71	19.19
LLaMA2 $_1$	2.090	**14.65**	**23.89**	13.44	**23.89**	2.222	11.11	19.21	12.92	19.21
LLaMA2 $_3$	2.393	5.36	16.09	9.97	16.09	2.386	5.93	16.41	13.74	16.41
LLaMA2 $_5$	2.417	4.15	15.48	2.39	15.48	2.417	4.15	15.48	2.39	15.48
Mistral $_0$	2.120	12.55	22.12	12.19	22.12	2.221	8.99	18.33	10.74	18.33
Mistral $_1$	2.114	12.78	22.29	12.42	22.29	2.269	8.57	18.26	10.42	18.26
Mistral $_3$	2.374	6.45	16.71	12.50	16.71	2.380	6.48	16.71	**15.36**	16.71
Mistral $_5$	1.739	6.29	18.58	5.98	18.58	**1.733**	6.28	18.58	7.25	18.58
Ensemble $_0$	2.111	14.00	23.21	13.33	23.21	2.277	10.00	19.31	13.78	19.31
Ensemble $_1$	2.102	14.25	23.52	13.74	23.52	2.269	10.39	19.50	13.91	19.50
Ensemble $_3$	2.374	6.45	16.71	12.50	16.71	2.405	4.77	15.79	9.95	15.79
Ensemble $_5$	2.080	7.65	15.17	5.13	15.17	2.402	4.21	15.48	2.43	15.48
					MBFC-2025					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.354	29.90	68.90	33.92	30.07	0.433	23.99	64.17	29.24	25.98
$BERT_{Base}$	0.276	37.28	75.59	40.17	36.76	0.287	37.62	75.20	**44.75**	36.56
$RoBERTa_{Base}$	0.260	**41.12**	75.98	42.71	**40.31**	0.280	34.60	74.41	35.21	34.64
$DistilBERT_{Base}$	**0.252**	38.03	**77.56**	43.23	37.00	0.276	35.61	76.38	40.72	35.10
$ALBERT_{Base}$	**0.252**	35.57	76.38	37.30	35.50	0.264	35.50	75.98	38.28	35.29
Ensemble	0.256	37.95	77.17	42.62	37.08	0.260	39.73	76.77	46.32	38.15
					LLMs					
LLaMA2 $_0$	1.541	14.57	23.84	13.95	23.84	**1.535**	**15.25**	**24.77**	14.45	**24.77**
LLaMA2 $_1$	2.516	2.97	12.90	1.67	12.90	2.516	2.95	12.90	1.66	12.90
LLaMA2 $_3$	2.516	2.95	12.90	1.66	12.90	2.516	2.95	12.90	1.66	12.90
LLaMA2 $_5$	2.516	2.95	12.90	1.66	12.90	2.516	2.95	12.90	1.66	12.90
Mistral $_0$	1.761	6.06	17.65	4.51	17.65	1.736	6.20	18.26	5.93	18.26
Mistral $_1$	2.419	4.15	13.71	3.26	13.71	2.483	3.01	12.90	1.71	12.90
Mistral $_3$	2.516	2.97	12.90	1.67	12.90	2.516	2.95	12.90	1.66	12.90
Mistral $_5$	1.782	1.32	8.06	0.72	8.06	1.790	1.33	8.06	0.73	8.06
Ensemble $_0$	1.705	12.65	17.03	**15.20**	17.03	1.770	11.99	18.89	13.68	18.89
Ensemble $_1$	2.467	3.03	12.90	1.72	12.90	2.516	2.95	12.90	1.66	12.90
Ensemble $_3$	2.516	2.95	12.90	1.66	12.90	2.516	2.95	12.90	1.66	12.90
Ensemble $_5$	2.112	3.86	10.48	2.40	10.48	2.459	3.10	12.90	1.76	12.90

Table 13: Evaluation results for *factuality* using hard and soft voting for each framework and in ensemble.

Models	Political Bias					Factuality				
	MAE	**Macro-F1**	**Accuracy**	**Precision**	**Recall**	**MAE**	**Macro-F1**	**Accuracy**	**Precision**	**Recall**
					ACL-2020					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.633	52.23	53.06	52.38	52.61	0.551	**36.92**	57.14	34.12	**40.33**
$BERT_{Base}$	0.673	41.62	48.98	45.66	47.31	**0.531**	30.20	**61.22**	37.23	36.11
$RoBERTa_{Base}$	**0.571**	**57.52**	**59.18**	**58.14**	**57.86**	0.571	24.79	59.18	19.73	33.33
$DistilBERT_{Base}$	0.653	24.63	38.78	29.55	35.40	0.571	24.79	59.18	19.73	33.33
$ALBERT_{Base}$	0.592	47.68	53.06	54.67	50.81	0.571	28.49	57.14	28.33	33.81
Ensemble HV	0.604	44.63	50.00	54.58	48.74	0.542	29.99	60.42	**36.96**	36.11
Ensemble SV	0.583	55.00	56.25	56.62	55.46	0.583	24.56	58.33	19.44	33.33
					LLMs					
LLaMA2	**0.568**	36.64	**45.95**	65.14	45.95	**1.536**	**15.25**	12.68	14.45	**24.77**
Mistral	0.746	40.65	45.74	61.17	45.74	1.737	6.07	**27.86**	5.93	18.27
Ensemble HV	0.640	**42.73**	48.02	**62.22**	**48.02**	1.706	12.65	21.62	**15.20**	17.03
Ensemble SV	0.655	41.11	46.19	59.86	46.19	1.706	12.17	20.80	14.63	16.39
					MBFC-2025					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.771	52.82	56.07	55.58	52.41	0.552	25.23	50.25	**27.02**	25.09
$BERT_{Base}$	0.645	60.86	62.15	63.86	60.42	**0.398**	27.04	63.68	25.49	28.80
$RoBERTa_{Base}$	0.692	58.98	60.75	59.29	59.28	0.413	**29.49**	63.68	45.47	**30.04**
$DistilBERT_{Base}$	0.766	54.14	58.88	61.82	54.01	**0.398**	27.45	**64.68**	26.07	29.18
$ALBERT_{Base}$	0.734	56.42	59.35	59.11	57.63	0.413	27.22	64.18	25.67	29.03
Ensemble HV	**0.623**	**63.19**	**64.15**	**65.33**	**62.86**	0.403	27.26	64.18	25.74	28.99
Ensemble SV	0.698	58.30	60.85	60.74	58.04	0.403	27.01	63.68	25.45	28.82
					LLMs					
LLaMA2	1.424	18.95	31.39	**27.31**	31.39	1.011	32.79	40.95	27.69	40.95
Mistral	1.418	9.06	14.86	21.88	14.86	**0.849**	**36.06**	**51.21**	28.12	**51.21**
Ensemble HV	**1.411**	16.01	24.15	24.30	24.15	0.927	34.72	46.21	27.83	46.21
Ensemble SV	1.426	**20.19**	**32.13**	27.07	**32.13**	0.872	36.01	49.24	**28.39**	49.24

Table 14: Evaluation results for *political bias* and *factuality* using frameworks independently and in ensembles using hard voting (HV) and soft voting (SV) at media-level Wikipedia datasets.

Models	Hard Voting					Soft Voting				
	MAE	**Macro-F1**	**Accuracy**	**Precision**	**Recall**	**MAE**	**Macro-F1**	**Accuracy**	**Precision**	**Recall**
					ACL-2020					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.490	47.36	53.06	76.11	51.12	0.490	50.00	55.10	69.36	53.50
$BERT_{Base}$	0.367	68.69	69.39	72.82	68.29	0.367	66.84	67.35	72.99	66.32
$RoBERTa_{Base}$	**0.224**	**81.47**	**81.63**	**83.61**	**81.20**	0.286	77.16	77.55	78.45	76.97
$DistilBERT_{Base}$	0.327	70.45	71.43	76.35	69.83	0.388	65.67	67.35	72.77	65.48
$ALBERT_{Base}$	0.551	56.63	61.22	60.04	58.65	0.449	62.02	65.31	67.86	62.99
Ensemble	0.375	67.76	68.75	71.20	67.36	0.354	70.02	70.83	73.64	69.51
					LLMs					
LLaMA2 $_0$	1.500	16.64	20.00	32.30	20.00	1.510	15.54	19.67	32.16	19.67
LLaMA2 $_1$	1.673	19.53	32.00	15.01	32.00	1.657	19.22	34.00	16.39	34.00
LLaMA2 $_3$	**1.233**	20.08	25.00	23.47	25.00	1.237	22.13	28.00	24.69	28.00
LLaMA2 $_5$	1.559	9.75	14.63	21.18	14.63	1.537	11.60	15.85	24.48	15.85
Mistral $_0$	1.440	10.73	24.67	19.74	24.67	1.400	12.77	24.33	19.53	24.33
Mistral $_1$	1.453	20.22	22.00	33.03	22.00	1.457	20.92	23.00	32.05	23.00
Mistral $_3$	1.493	19.66	21.00	36.86	21.00	1.483	18.35	20.00	**36.18**	20.00
Mistral $_5$	1.375	21.43	23.17	36.21	23.17	1.358	20.07	21.95	36.31	21.95
Ensemble $_0$	1.460	16.61	24.00	25.24	24.00	1.490	17.51	21.67	31.73	21.67
Ensemble $_1$	1.590	21.38	27.00	20.62	27.00	1.350	**28.87**	**35.00**	26.08	**35.00**
Ensemble $_3$	1.363	20.95	23.00	27.97	23.00	1.307	21.86	25.00	26.61	25.00
Ensemble $_5$	1.452	17.21	19.51	37.04	19.51	1.537	11.60	15.85	24.48	15.85
					MBFC-2025					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.712	48.47	58.02	64.04	49.03	0.708	47.63	57.55	63.41	48.28
$BERT_{Base}$	0.566	66.21	69.81	69.04	66.02	0.491	69.19	72.17	70.38	68.70
$RoBERTa_{Base}$	0.500	66.54	70.28	67.68	66.43	0.505	67.03	70.75	68.19	66.75
$DistilBERT_{Base}$	0.519	**67.13**	**71.23**	**68.81**	**67.43**	0.495	66.03	70.75	67.99	66.26
$ALBERT_{Base}$	0.500	68.23	71.23	69.59	68.60	0.538	68.32	71.70	69.76	68.99
Ensemble	**0.448**	70.23	73.58	72.09	70.37	**0.425**	**72.41**	**75.47**	**73.69**	**72.36**
					LLMs					
LLaMA2 $_0$	1.500	16.64	19.00	**40.47**	19.00	1.510	15.54	18.00	40.15	18.00
LLaMA2 $_1$	1.673	18.22	29.67	14.71	29.67	1.657	19.22	30.67	15.61	30.67
LLaMA2 $_3$	**1.233**	**24.18**	**33.00**	25.19	**33.00**	1.237	24.12	33.00	24.76	33.00
LLaMA2 $_5$	1.559	4.23	12.37	7.32	12.37	1.549	5.34	13.38	8.85	13.38
Mistral $_0$	1.440	10.73	22.00	7.18	22.00	1.400	10.39	22.00	6.82	22.00
Mistral $_1$	1.453	19.23	23.67	29.66	23.67	1.457	19.81	24.33	29.74	24.33
Mistral $_3$	1.493	15.89	20.33	**31.69**	20.33	1.483	16.79	21.00	33.03	21.00
Mistral $_5$	**1.375**	19.93	25.75	26.94	25.75	**1.358**	20.32	26.76	26.80	26.76
Ensemble $_0$	1.460	15.18	19.00	24.43	19.00	1.500	18.35	20.00	36.18	20.00
Ensemble $_1$	1.520	20.59	28.67	17.69	28.67	**1.320**	**26.61**	**36.67**	22.07	**36.67**
Ensemble $_3$	1.363	22.87	29.00	26.56	29.00	1.307	24.57	32.33	28.14	32.33
Ensemble $_5$	1.452	15.26	20.74	27.40	20.74	1.498	10.71	18.06	11.41	18.06

Table 15: Evaluation results for *political bias* using hard and soft voting for each framework and in the ensemble.

Models	Hard Voting					Soft Voting				
	MAE	**Macro-F1**	**Accuracy**	**Precision**	**Recall**	**MAE**	**Macro-F1**	**Accuracy**	**Precision**	**Recall**
					ACL-2020					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.571	24.79	59.18	19.73	33.33	0.571	24.79	59.18	19.73	33.33
$BERT_{Base}$	0.531	25.44	59.18	20.57	33.33	0.551	25.11	59.18	20.14	33.33
$RoBERTa_{Base}$	0.510	30.22	61.22	32.13	36.11	0.510	30.22	61.22	32.13	36.11
$DistilBERT_{Base}$	0.551	30.24	61.22	53.47	36.11	0.551	30.24	61.22	53.47	36.11
$ALBERT_{Base}$	**0.469**	**42.42**	**65.31**	**77.04**	**43.06**	0.510	34.67	63.27	43.24	38.89
Ensemble	0.500	30.89	62.50	32.59	36.36	0.521	30.91	62.50	37.68	36.36
					LLMs					
LLaMA2 $_0$	1.990	10.52	20.00	7.83	20.00	1.950	11.59	22.00	8.46	22.00
LLaMA2 $_1$	2.470	3.74	11.00	21.01	11.00	2.490	1.82	10.00	1.00	10.00
LLaMA2 $_3$	1.430	20.08	25.00	23.47	25.00	1.390	**22.13**	**28.00**	24.69	**28.00**
LLaMA2 $_5$	2.484	1.87	10.10	1.03	10.10	2.485	1.87	10.10	1.03	10.10
Mistral $_0$	2.490	1.82	10.00	1.00	10.00	2.490	1.82	10.00	1.00	10.00
Mistral $_1$	2.470	3.74	11.00	21.01	11.00	2.470	3.74	11.00	21.01	11.00
Mistral $_3$	1.490	19.66	21.00	36.86	21.00	1.500	18.35	20.00	**36.18**	20.00
Mistral $_5$	2.313	6.28	14.14	5.30	14.14	2.313	6.28	14.14	5.30	14.14
Ensemble $_0$	2.310	6.47	12.13	4.36	12.13	2.050	9.83	18.00	7.90	18.00
Ensemble $_1$	2.380	4.56	12.12	4.36	12.12	2.470	3.74	11.00	21.01	11.00
Ensemble $_3$	1.480	20.94	23.00	27.97	23.00	**1.440**	21.86	25.00	26.61	25.00
Ensemble $_5$	2.484	1.87	10.10	1.03	10.10	2.505	1.85	10.10	1.02	10.10
					MBFC-2025					
					PLMs					
$SVM_{TF\text{-}IDF}$	0.448	25.88	62.69	27.18	28.55	0.443	26.20	63.18	27.14	28.76
$BERT_{Base}$	**0.303**	33.07	72.14	38.94	33.82	0.328	32.32	70.15	48.02	32.92
$RoBERTa_{Base}$	0.313	**41.58**	71.64	68.58	38.63	0.323	30.12	71.14	28.47	32.22
$DistilBERT_{Base}$	0.308	33.31	**72.64**	**49.01**	**34.08**	0.318	32.92	71.64	48.59	33.62
$ALBERT_{Base}$	0.343	31.93	69.65	37.98	32.75	0.333	32.19	70.15	38.13	32.95
Ensemble	0.313	33.07	72.14	48.84	33.88	0.328	32.28	70.15	47.98	32.95
					LLMs					
LLaMA2 $_0$	1.953	10.65	19.67	7.99	19.67	**1.947**	**11.25**	**20.67**	8.46	**20.67**
LLaMA2 $_1$	2.400	3.10	12.00	8.15	12.00	2.413	2.45	11.67	1.37	11.67
LLaMA2 $_3$	2.413	2.45	11.67	1.37	11.67	2.413	2.45	11.67	1.37	11.67
LLaMA2 $_5$	2.426	2.47	11.75	1.38	11.75	2.426	2.47	11.75	1.38	11.75
Mistral $_0$	2.393	2.47	11.67	1.38	11.67	2.400	2.46	11.67	1.37	11.67
Mistral $_1$	2.400	4.37	12.67	21.71	12.67	2.400	4.37	12.67	**21.71**	12.67
Mistral $_3$	2.420	2.44	11.67	1.36	11.67	2.420	2.44	11.67	1.36	11.67
Mistral $_5$	2.332	6.23	13.76	10.22	13.76	2.332	6.23	13.76	10.22	13.76
Ensemble $_0$	2.193	6.66	14.67	6.38	14.67	2.053	10.26	18.33	8.51	18.33
Ensemble $_1$	2.400	3.73	12.33	14.93	12.33	2.413	3.10	12.00	8.15	12.00
Ensemble $_3$	2.413	2.44	11.67	1.37	11.67	2.420	2.44	11.67	1.36	11.67
Ensemble $_5$	2.352	4.25	12.75	6.92	12.75	2.426	2.47	11.75	1.38	11.75

Table 16: Evaluation results for *factuality* using hard and soft voting for each framework and in the ensemble.