

Read Between the Lines: A Benchmark for Uncovering Political Bias in Bangla News Articles

Nusrat Jahan Lia
University of Dhaka
bsse1306@iit.du.ac.bd

Shubhashis Roy Dipta
University of Maryland,
Baltimore County
sroydip1@umbc.edu

Abdullah Khan Zehady
Cisco Systems
azehady@cisco.com

Naymul Islam
BanglaLLM
naymul504@gmail.com

Madhusodan Chakraborty
Maharishi International University
opuchakraborty@gmail.com

Abdullah Al Wasif
Unityflow AI
wasif@unityflow.ai

Abstract

Detecting media bias is crucial, specifically in the South Asian region. Despite this, annotated datasets and computational studies for Bangla political bias research remain scarce. Crucially because, political stance detection in Bangla news requires understanding of linguistic cues, cultural context, subtle biases, rhetorical strategies, code-switching, implicit sentiment, and socio-political background. To address this, we introduce BanglaBias, the first benchmark dataset of 200 politically significant and highly debated Bangla news articles, labeled for government-leaning, government-critique, and neutral stances, alongside diagnostic analyses for evaluating large language models (LLMs). Our comprehensive evaluation of 28 proprietary and open-source LLMs shows strong performance in detecting government-critique content (F1 up to 0.83) but substantial difficulty with neutral articles (F1 as low as 0.00). Models also tend to over-predict government-leaning stances, often misinterpreting ambiguous narratives. BanglaBias and its associated diagnostics provide a foundation for advancing stance detection in Bangla media research and offer insights for improving LLM performance in low-resource languages.¹

1 Introduction

News media often reflect subtle political framing and bias, shaping public opinion and behavior (Fuhat and Wahab, 2024). While extensive research has addressed this problem in English by creating rich datasets for sentiment analysis (Elbounani et al., 2025; Abercrombie and Batista-Navarro, 2020), stance detection (Rostami et al., 2025; Khiabani and Zubiaga, 2025), hate speech

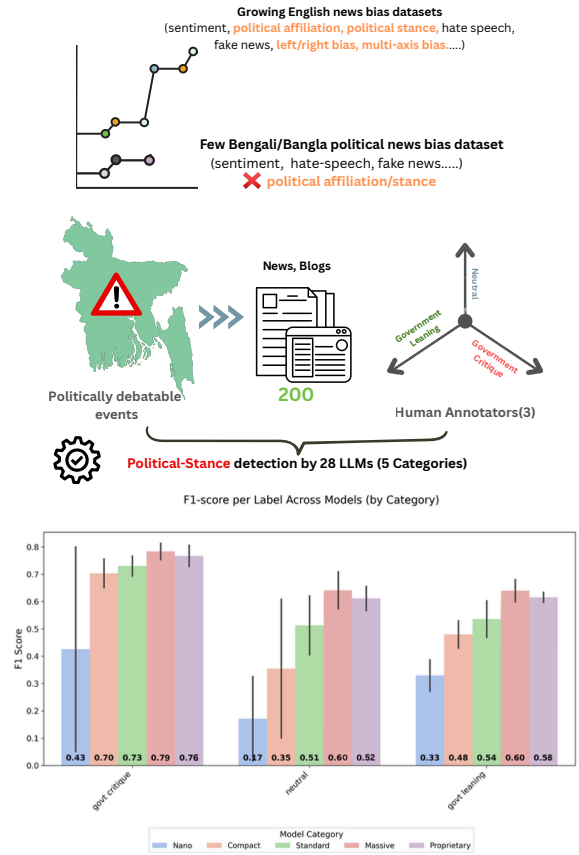


Figure 1: Overview of political stance detection study (**Growing Resources for English vs. Lack of Bangla Resource Availability**). We introduce a benchmark of 200 news articles (on politically debatable events) annotated into Government Leaning, Critique, and Neutral labels. We then evaluate performance of 28 LLMs in detecting political stance in Bengali. Performance improves with model size, with Massive and Proprietary models achieving highest F1-scores, but neutral detection remains weak. Bars for Nano models and Neutral label show **noticeably larger error ranges across models, indicating unstable performance**.

¹<https://nusrat-lia.github.io/BanglaBias/>

identification (Davidson et al., 2017), and political bias classification (Jones, 2024), the same cannot be said for many other languages. As Fig. 1 illustrates, Bangla news media research faces a critical gap: despite clear evidence of political framing in coverage (Islam, 2016), there exists a scarcity of annotated datasets specifically designed to detect political stance and government affiliation bias. Our evaluations in Fig. 1 revealed that while massive or proprietary models achieve strong performance on critique stance detection, neutral stance remains challenging, and Nano models showing large error bars underscores unstable performance in low-resource scenarios.

This disparity is particularly concerning given the widespread nature of political bias in Bangla news. Existing studies (both qualitative and computational) provide clear evidence that Bangla news content exhibits partisan frames that favor ruling-party narratives (Islam, 2016; Al-Zaman and No-man, 2024). For instance, two news outlets covering the same political event, such as a government policy announcement, may frame it as either a progressive reform or an authoritarian overreach, despite reporting the same factual details (Vallejo et al., 2023). Such framing differences are not merely stylistic; they represent systematic *ideological bias* – a partisan slant away from neutral reporting (McQuail and Deuze, 2020).

Computational studies have begun to tackle this with machine learning, e.g., Bangla BERT (Bhattacharjee et al., 2021) for hyper-partisanship (Mehadi Hasan et al., 2025) and innovative LLM credibility bias (Prana and Islam, 2025). However, these efforts are hampered by the lack of comprehensive and well-annotated datasets that can evaluate robust detection of political positions, specifically for the Bangladeshi socio-political landscape. Audience studies further highlight how perceived slant and censorship are reshaping Bangla news consumption (Islam et al., 2025).

To address this gap, we begin with collecting politically debatable events sourced from diverse news outlets and blogs. We then curated a dataset of 200 samples annotated by three native speakers for government stance (Government Leaning, Government Critique, or Neutral). BanglaBias addresses the challenges of detecting political stance in Bangla news, as this is far more challenging than in English due to limited prior corpora, nuanced political language, frequent code-switching with English terms, and context-dependent rhetorical

styles that make bias subtle and culturally embedded. We then evaluate reasoning-based political stance detection across 28 large language models (proprietary and open-source) to assess their effectiveness for this task.

To summarize, our work revolves around three key contributions:

- We introduce the first-ever benchmark dataset for political stance detection in Bangla with comprehensive metadata and a leaderboard. The dataset provides contextual information to study how political narratives are constructed in Bangladeshi media and enables the development and evaluation of stance detection models in such a low-resource setting.
- We demonstrate multiple LLMs’ effectiveness in detecting political bias in low-resource settings, with larger models achieving strong performance on Government Critique (F1=0.78) and Neutral (F1=0.61) classifications.
- We provide a systematic analysis that reveals important performance differences between different sizes of models as well as their biases and error tendencies.

2 Relevant Work

With the growing body of research on framing in NLP (Card et al., 2015; Fan et al., 2019; Baly et al., 2020; Ziemis and Yang, 2021), and LLM prompting strategies (Brown et al., 2020; Achiam et al., 2023; Touvron et al., 2023) to detect stances, recent studies have explored media bias through multiple computational approaches, as we discuss in the following sections.

2.1 Framing & Political Bias in News Media

Entman (2007) defines framing as the process of shaping narratives to promote specific interpretations, which then primes audiences to think in particular ways. Computational NLP work has built on these ideas by annotating and classifying news frames and biases (Card et al., 2015; Guida et al., 2025). Early efforts include the Media Frames Corpus (Card et al., 2015) and its multilingual extension (Piskorski et al., 2023), while others focus on ideological bias, such as BASIL (Fan et al., 2019), AllSides (Baly et al., 2020), and smaller resources on regional or issue-specific perspectives (Lin et al., 2006; Chen et al., 2018; Ziemis and Yang, 2021).

Such datasets do not capture the socio-political nuances of Bangladeshi cultural context, and there exists no such Bengali counterpart yet.

In practice, most NLP work treats media bias as a classification problem: one common approach is single-label classification of an article’s ideology. For instance, [Recasens et al. \(2013\)](#) derived word-level ideological features; [Spinde et al. \(2021\)](#) identified biased words via embedding distances between left- and right-leaning. Transformer-based models (e.g., BERT variants) now dominate framing tasks ([Liu et al., 2019](#); [Akyürek et al., 2020](#); [Piskorski et al., 2023](#)). [Mehadi Hasan et al. \(2025\)](#) presented a Bangla BERT model fine-tuned to identify hyperpartisan (extremely biased) news articles, achieving 95.7% accuracy, using a semi-supervised approach. The authors emphasized Bangla’s “low-resource” status – few annotated corpora, lexicons or pretrained models, and therefore relied on clever workarounds, i.e., machine translation (MT) for data augmentation. While MT is a practical approach, it fails to identify the socio-political facts that are specific to the region.

2.2 LLMs for Political & Subjective Analysis

Large pretrained language models have become popular for political stance and bias tasks via prompt or instruction-based methods. Researchers have evaluated models like GPT-3/GPT-4 and open alternatives (Falcon, LLaMA, Mistral, etc.) on stance detection and bias classification ([Faulborn et al., 2025](#); [Ng et al., 2025](#)). For example, [Ng et al. \(2025\)](#) tested GPT-3.5 and seven open-source LLMs on three stance datasets (SemEval-2016 tweets ([Mohammad et al., 2016](#)), Elections-2016 ([Sobhani et al., 2017](#)) tweets, and the BASIL ([Fan et al., 2019](#)) news articles); they tried different prompting schemes (task description, context, chain-of-thought) and found reasoning based prompting scheme demonstrating the best performance in stance classification.

[Sucu et al. \(2025\)](#) showed that adding user-context information to prompts can boost performance: in a political forum stance task, contextual prompting improved accuracy by 17.5% to 38.5% over baseline. [Lihammer \(2023\)](#) showed GPT-3.5 and GPT-4 being able to reproduce political viewpoints. When tested in Bangladeshi context, most LLMs favored the left-leaning sources, giving higher credibility ratings ([Prama and Islam, 2025](#)). Instruction-tuned LMs can perform stance/bias tasks in a zero-shot or few-shot man-

ner, but their outputs must be carefully validated as they may favor certain viewpoints and result in AI-induced bias in evaluating news sources, more significantly in low-resource languages like Bangla ([Ng et al., 2025](#); [Prama and Islam, 2025](#)).

3 Creating BanglaBias

Prior research on Bangla social discourse classification ([Haider et al., 2024](#)) highlighted the challenges of transliteration noise, subtle category distinctions, and embedding reliability. Beyond these technical difficulties, a major bottleneck remains – the scarcity of high-quality, publicly available, politically nuanced Bangla datasets, which limits the development and evaluation of models for political stance detection and bias analysis. While translation or synthetic dataset generation has shown promise in domains like mathematics ([Toshniwal et al., 2024](#)), instruction following ([Kim and Park, 2024](#)), or general language tasks ([Liu et al., 2024](#)), these approaches are often ineffective for political stance detection and media bias analysis in Bangla as socio-political contexts are deeply tied to cultural, historical, and linguistic nuances that are lost or distorted in translation. Building on the research gap, we developed a systematic pipeline to create the first annotated Bangla political stance detection benchmark dataset.

For collecting the articles and human annotations, we followed these steps:

Event Selection and Categorization: We identified 46 socio-political events (full list in App. D) that generated significant public controversy or divergent viewpoints across the political spectrum. Our selection criteria prioritized events that: (1) received substantial media coverage across multiple outlets, (2) elicited clear government support or criticism in public discourse, and (3) represented diverse policy domains to ensure broad coverage of political stance patterns.

The temporal distribution spans Bangladesh’s political landscape during the years 2013–2025. The early years (2013–2015) witnessed industrial disasters, the Shahbagh and Hefazat movements, and the contested 2014 election. The middle phase (2016–2019) was marked by the Holey Artisan attack, the Rohingya refugee influx, student protests on quota reform and road safety, and the 2018 election. The pandemic years (2020–2021) brought COVID-19, garment wage unrest, anti-rape protests, and growing dissent. The recent phase (2022–2025) includes

fuel price hikes, the Padma Bridge opening, Rampal power plant launches, BNP–Awami League mega rallies, the boycotted 12th election, and violent quota protests with internet shutdowns and curfews.

Crawling: We implemented a crawler to collect news articles for each event title based on the year of occurrence and saved results in HTML format. If no relevant articles were retrieved for the target year, the search was extended to the following 5-6 years to capture delayed or retrospective coverage.

Parsing and normalization: The HTML files were parsed and normalized to retain only the relevant data points, excluding advertisements, navigation menus, and unrelated web content.

Article selection: For each event, at least one and up to five representative articles were selected. We prioritized articles from established mainstream outlets including Prothom Alo, The Daily Star (Bangla), Jugantor, and Kaler Kantho, alongside politically diverse sources spanning the ideological spectrum.

Human Annotation: News items were annotated for political stance using a three-category framework: *Government Leaning* (content that supports or favors government positions), *Government Critique* (content that criticizes government actions), and *Neutral* (balanced reporting without clear partisan positioning). We focused on identifying implicit stance indicators beyond explicit political statements, including source selection, framing choices, and political landscape analysis of each article’s timeline and contextual emphasis.

Two annotators performed the initial annotation in parallel. Disagreements between these two annotators were resolved by a third annotator (Annotator 3), who acted as the adjudicator; the adjudicated labels are reported as final label (see annotator information in App. F).

Annotator 1 and Annotator 2 achieved 73.5% agreement (Cohen’s $\kappa = 0.574$), indicating moderate inter-rater reliability. Most of the disagreements arose from neutral vs government critique and some neutral vs government leaning confusions. Disagreements were resolved by a third annotator (adjudicator) after mutual discussion and reanalysis of the political affiliations of the respective news media. Annotators primarily adhered to the stance detection decision flow outlined in App. H.

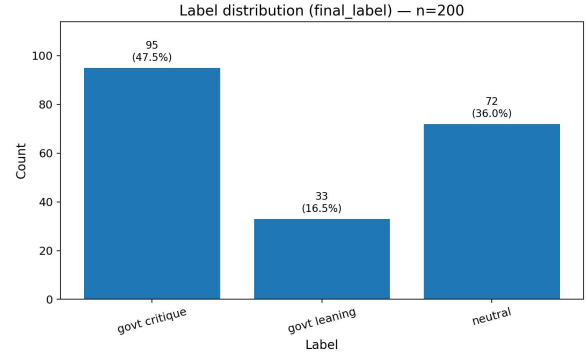


Figure 2: Distribution of three classes across the dataset: 95 Govt. Critique (47.5%), 72 Neutral (36.0%), and 33 Govt. Leaning (16.5%).

Our event-driven design captures politically significant moments where public opinion and media polarization are most visible, considering nuanced stance signals across politically diverse outlets. This also ensures quality news dataset while decreasing the chances of noisy data. **Our news collection pipeline is reusable and extensible to any low- or high-resource language.** This allows future researchers to expand the dataset with new events or apply it to other political contexts.

3.1 Dataset Statistics

In total there are 200 data samples and 8 key information; every item has a unique *ID*, a *News Body*, a *Headline*, a *Source link* and the *Date*. The *Event* field contains 46 distinct events, and *News Corpora Name* lists 54 distinct sources with the largest contributors being BBC Bangla (N=37), Prothom Alo (N=20), Jugantor (N=17), DW (N=11) and Bangla Tribune (N=11) (full details reported in Table 4). As Fig. 2 presents, the final label has three classes (govt. critique: 95, Neutral: 72, govt. leaning: 33). Because nearly half the dataset is labeled “govt. critique” while “govt. leaning” is underrepresented ($\approx 16.5\%$), we recommend treating class imbalance explicitly during evaluation.

4 Experimental Setup

We benchmark the performance of several state-of-the-art large language models (LLMs) on the political stance detection using our benchmark dataset of 200 Bangla articles annotated with one of three stances: Government Leaning, Government Critique, or Neutral.

		<i>Govt. Critique</i>			<i>Neutral</i>			<i>Govt. Leaning</i>			<i>Weighted Avg.</i>		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Nano	Qwen3-1.7B	0.71	0.69	0.70	0.54	0.21	0.30	0.28	0.67	0.39	0.58	0.52	0.51
	Qwen3-0.6B	0.58	0.57	0.57	0.35	0.15	0.21	0.22	0.52	0.31	0.44	0.41	0.40
Compact	TigerLLM-9B	0.74	0.78	0.76	0.62	0.65	0.64	0.62	0.45	0.53	0.68	0.68	0.68
	Qwen3-8B	0.75	0.73	0.74	0.62	0.42	0.50	0.42	0.76	0.54	0.65	0.62	0.62
Standard	Qwen3-32B	0.82	0.68	0.75	0.55	0.72	0.63	0.70	0.58	0.63	0.71	0.68	0.68
	GPT-OSS-20B	0.75	0.72	0.73	0.55	0.60	0.57	0.55	0.52	0.53	0.64	0.64	0.64
Massive	Qwen3-235B-I	0.78	0.86	0.82	0.71	0.62	0.67	0.66	0.64	0.65	0.74	0.74	0.74
	Llama3.3-70B	0.77	0.91	0.83	0.80	0.50	0.62	0.57	0.76	0.65	0.75	0.73	0.73
Proprietary	Claude-Sonnet-4	0.90	0.73	0.80	0.63	0.69	0.66	0.52	0.70	0.60	0.74	0.71	0.72
	Gemini-2.5-Pro	0.79	0.77	0.78	0.61	0.61	0.61	0.56	0.61	0.58	0.69	0.69	0.69

Table 1: Performance metrics (Precision, Recall, F1) for the top two models per category, ranked by weighted F1.

4.1 Evaluation Metrics

Our primary evaluation frames political stance detection as a TERNARY CLASSIFICATION task. Each data point consists of an article text and a ground-truth label (Government Leaning, Neutral, or Government Critique). Given the article, models must predict the stance. We evaluate models on this task using standard metrics including accuracy, precision, recall, and weighted F1 score.

We conduct evaluations in a reasoning-based detection setting. Following Ng et al. (2025), we prompted models to classify the article and provide a brief reasoning. This framework supports additional qualitative analyses to reveal performance differences across model sizes:

- (1) **Per-Label Performance** assesses F1 scores for each stance category to identify imbalances in handling Government Leaning, Neutral, or Government Critique articles.
- (2) **Confusion Analysis** uses confusion matrices (aggregated by model category) to highlight common misclassification patterns, such as confusing Neutral with Government Leaning.
- (3) **Bias Tendency** evaluates prediction distributions via radar plots, comparing them to the ground-truth distribution to detect systematic biases toward certain stances.
- (4) **Misclassification Analysis** examines incorrect predictions alongside model-generated reasons to uncover qualitative error patterns, such as over-reliance on specific keywords or failure to detect nuance.

4.2 Models

We evaluate a diverse set of 28 language models, categorized by parameter size: Nano (<2B parameters), Compact (2B–10B), Standard (10B–40B), Massive (>40B), and Proprietary (size undisclosed). This selection includes both open-source and closed-source models to ensure comprehensive coverage.

- **Nano:** Qwen2.5-0.5B, Qwen3-0.6B, Qwen3-1.7B.
- **Compact:** Qwen2.5-3B, Qwen3-4B, Qwen3-4B-I, Qwen3-4B-T, Qwen3-8B, Llama3.1-8B, TigerLLM-9B.
- **Standard:** Qwen2.5-14B, Qwen3-14B, Qwen3-30B-I, Qwen3-30B-T, Qwen3-32B, GPT-OSS-20B.
- **Massive:** Llama3.3-70B, GPT-OSS-120B, Qwen3-235B-I, Qwen3-235B-T, GLM-4.5, DeepSeek-V3.1, DeepSeek-R1.
- **Proprietary:** Claude-Sonnet-4, Grok-4, Gemini-2.5-Pro, GPT-4.1-Mini, GPT-5-Mini.

Models are prompted to generate both a stance label (Decision: D) and reasoning (R) ($R \rightarrow D$). Full model ids can be found in App. E. The evaluated models include families such as Qwen (Bai et al., 2023), LLaMA (Touvron et al., 2023), GLM (Zeng et al., 2025), DeepSeek (Bi et al., 2024), GPT (Achiam et al., 2023), Claude (Caruccio et al., 2024), and Gemini (Ng et al., 2025).

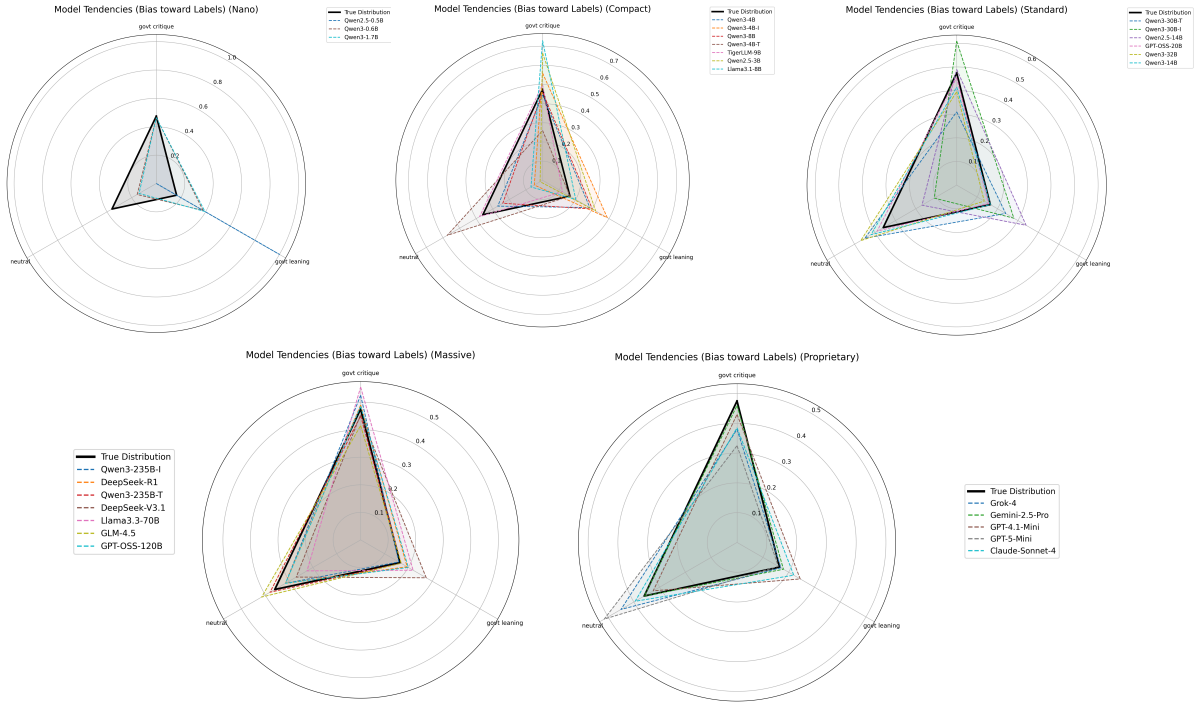


Figure 3: Radar plots showing tendencies of models (per-category) to favor particular labels, relative to the true distribution. The black polygon in each radar plot denotes the true distribution of labels and serves as the baseline.

5 Results & Analysis

5.1 Performance Metrics Analysis

Table 1 presents the performance of the top two models from each category (ranked by weighted average F1 score) in political stance detection task. Results for all 28 models can be found in Table 2.

★ Finding 1: Neutral Category Exposes Model Weaknesses

Most models struggle with neutral content, with F1 scores as low as 0.00 for Qwen2.5-0.5B and 0.16 for Llama3.1-8B due to poor recall (e.g., 0.10 for Llama3.1-8B). Even massive models like Qwen3-235B-T only reach 0.68 (F1). The consistent struggle across model sizes points to potential underrepresentation of Bangla neutral samples and insufficient learning of contextual cues that distinguish neutral stances from biased ones. The challenge is particularly evident in articles requiring balanced interpretation of government actions without explicit sentiment, indicating a gap in models' ability to handle nuanced, non-polarized Bangla narratives (see Table 2).

★ Finding 2: Government Critique Stance is the Easiest One to Classify

Models excel at detecting government critique stance, with top performers Llama3.3-70B and Qwen3-235B-I achieving F1 scores of 0.83 and 0.82. We hypothesize that, critical contents contain distinct linguistic cues such as strong negative sentiment and keywords (e.g., Government negligence, fraud, Repressive policy, Negligence). This strength highlights a potential bias in model design, as the same models struggle with neutral content, suggesting an over-reliance on sentiment cues rather than nuanced contextual understanding.

★ Finding 3: Bias Towards Government-Leaning Stance

In government-leaning stance prediction, massive models like Qwen3-235B-I and Llama3.3-70B achieve decent F1 scores, but recall often exceeds precision (Table 2). This indicates a tendency to over-predict supportive stances. This imbalance suggests that models are overly sensitive to cues that might indicate government support, leading to false positives. This over-prediction risks am-

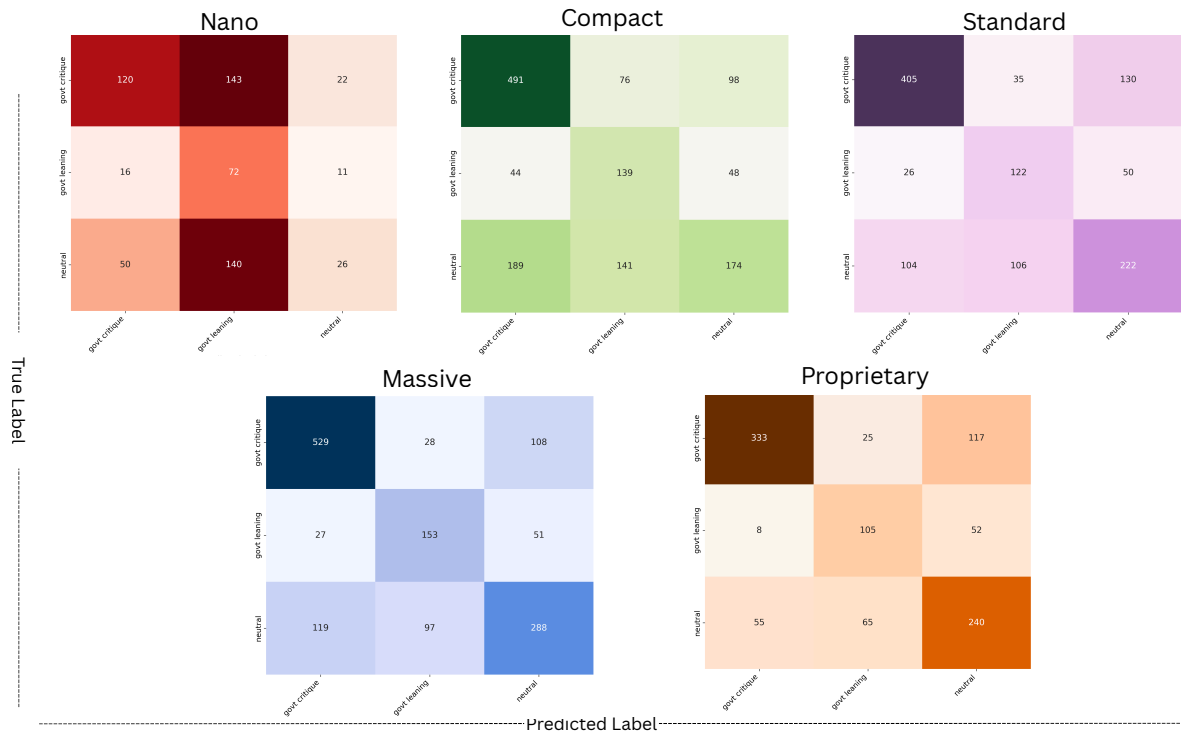


Figure 4: Aggregated Confusion Heatmap over five categories of models: Nano (3 models), Compact (7 models), Standard (6 models), Massive (7 models) and Proprietary (5 models).

plifying perceived government support in media analysis, potentially skewing automated content moderation or sentiment analysis applications.

★ **Finding 4:** Smaller Models can be Efficient (to some extent)

Although large-scale models (Massive and Proprietary) demonstrate clear dominance in classification tasks, standard models like Qwen3-32B and even smaller TigerLLM (Raihan and Zampieri, 2025) show competitive results in detecting stances, particularly government critique and neutral. This efficiency suggests that smaller models can effectively handle key linguistic features, such as critical keywords or neutral tone indicators, without requiring the computational resources of their larger counterparts. The strong performance of TigerLLM, in particular, highlights the potential of small domain-optimized models tailored for Bangla text, which benefit from focused training on regional linguistic patterns. In low-resource settings, where computational constraints are a significant concern, such resource-efficient models can be further optimized to perform critical stance detection tasks.

5.2 Bias Tendency

Fig. 3 illustrates models' tendencies across Nano, Compact, Standard, Massive, and Proprietary categories to favor particular labels, depicting biases in understanding Bangla linguistic variations. Nano models like Qwen2.5-0.5B, Qwen3-1.7B display extreme bias toward government leaning stance. Compact models vary in biases, with TigerLLM-9B showing decent alignment. Qwen3-4B-T leans towards neutral more often, whereas Llama3.1-8B mostly mistakes neutral articles as critique stances. Standard models like Qwen2.5-14B show bias towards government leaning stance prediction, while Qwen3-30B-I struggles with over-predicting critique label. Massive models including Llama3.3-70B and Qwen3-235B-I cluster closer to the true distribution with minimal deviations. Proprietary models like Gemini-2.5-Pro and GPT-4-Mini align well overall, but GPT-5-mini leans more towards neutral label.

But Where the Bias Lies? The confusion heatmaps in Fig. 4 reveal that smaller model categories like Nano and Compact frequently confuse neutral labels with government-leaning or critique, as seen in Nano's 140 neutral instances misclassified as government-leaning and Compact's 189

neutrals mistaken for government-critique. This indicates a bias toward polarized predictions due to limited capacity. As model scale increases, such as in Standard and Massive categories, mispredictions decrease, with Massive models showing balanced performance but still confusing neutral with government-critique. Proprietary models minimize cross-polarity errors (only 25 government-critique predicted as government leaning), yet persist in confusing neutral with polar stances (55 and 65 instances).

5.3 Prompting Strategy Comparison:

To further investigate the impact of prompting on stance detection, we compare zero-shot (no example) and few-shot (multiple examples) prompting strategies across selected models, focusing on those with varying sizes (e.g., 1.7B to 70B parameters).

In subjective tasks like political bias classification in Bangla news articles, few-shot prompting can underperform compared to zero-shot, particularly in larger models (see App. I), due to risks of overfitting to example patterns, such as over-emphasizing neutrality if examples are imbalanced or create conservative heuristics that dismiss implicit biases as “factual.” This leads to systematic errors like defaulting to “Neutral” labels, as seen in aggregate metrics where few-shot Weighted F1 scores lag (e.g., 0.57–0.67 vs. zero-shot’s 0.68–0.73). However, few-shot may still edge out in smaller models (e.g., 1.7B) needing more examples for calibration. This highlights that prompting-strategy efficacy depends on model scale, multiple patterns of framing and dataset nuances like implicit political tones.

5.4 Error Analysis

Table 3 shows some example error cases with the LLMs’ reasoning. Error analysis of model predictions reveals three main failure modes.

Content-balance Ambiguity: Many misclassified articles are fact-based, presenting both government and opposition, or including technical detail that lacks an evaluative authorial stance. Such articles require understanding of external political events. For such items, the ground-truth labels often force a single polar stance, but models predict neutral as they do not detect any favoring language.

In such instances, models often reason that the article presents both parties and hence they conclude it to be neutral. Such reasoning misses the

polar stance as it fails to incorporate understanding of any external entity or situational complexity.

For example, as shown in Table 3, in Article ID 110 (predicted neutral by Qwen3-32B, true label: govt. critique), the model reasoned that the article reports conflicting claims between protesters and the government-aligned student organization without taking a clear stance, presenting both accusations and denials. However, this overlooks the subtle emphasis on critiques through the framing of events, and shows how balanced presentation can mask underlying bias when external context (e.g., ongoing political tensions) is not considered.

Lexical-cue Over-reliance: When articles contain explicit praise (honorifics, laudatory framing), models predict government leaning stance even if the article was neutral. This stems from the fact that in Bangla articles, the mention of any leader name often comes with a laudatory term (“Respected Minister”), even though the articles are factual. Articles praising government initiatives when the actions were actually praise worthy, were classified as government leaning stance, despite it being neutral.

We hypothesize that as the article praises such persona, it must mean a favoring stance. Predicting such article stance is much harder as it requires world knowledge and context outside the immediate text (Burnham, 2025).

Examples include Article 98 (predicted govt. leaning by DeepSeek-V3.1, true label: neutral) (see Table 3), where the focus on a government lockdown was seen as supportive due to proactive framing, ignoring the fact based neutral trait. Similarly, in Article 148 (predicted govt. leaning by Qwen3-235B-I, true label: neutral), models reasoning highlighted government quotes against BNP, ignoring the objective situational context.

Selective Perspective Emphasis: A third common failure mode involves models amplifying one side’s narrative due to disproportionate quoting or structural prominence, leading to wrong predictions in otherwise balanced articles. This occurs particularly in dynamic political contexts where opposition voices are detailed extensively, even if countered, causing models to infer critique stance.

Similarly, events framed around protests or right-abuses push models toward government critique stance prediction, despite largely factual reporting. These patterns suggest models’ lack of understanding of speaker-aware representation and dependency on generic sentiment cues.

An illustrative case is Article 25 (predicted govt. critique by Qwen3-4B-T, true label: neutral) (see Table 3, where the reasoning emphasized the opposition’s (BNP) detailed accusations against the government, including human rights violations, despite the news being just a factual representation of opinions. The model failed to recognize the fact based patterns, treating vivid critical language as dominant bias.

6 Conclusion

Detecting political stance in Bangla news media requires nuanced understanding of linguistic cues, cultural context, and subtle biases embedded in reporting. This paper introduces a novel benchmark dataset, BanglaBias, designed to evaluate the political stance detection capabilities of computational approaches in media bias research, which can be extended to any low-resource setting. Our findings reveal that while most LLMs excel at identifying government-critical content, they struggle significantly with neutral content due to inadequate handling of context-dependent rhetorical styles, pointing to an under-representation of extensive samples in models’ training data. Our analyses indicate a tendency to misinterpret ambiguous content as government-leaning. Significant performance gains by domain-optimized LLMs further fortify our findings on the need for high-quality, culturally grounded datasets and targeted fine-tuning to improve stance detection in low-resource, politically nuanced contexts. Extending BanglaBias framework to other underrepresented languages could further democratize computational media analysis and sets the stage for more equitable, accurate, and culturally informed bias detection systems globally.

Limitations

Our study evaluates a diverse set of 28 large language models (LLMs) for political stance detection in Bangla news, covering both proprietary and open-source architectures. However, we acknowledge the rapid development of newer models or specialized architectures. We plan to extend our evaluation to incorporate emerging models and domain-specific adaptations to better capture Bangla-specific nuances.

BanglaBias, while carefully curated to represent government-leaning, government-critique, and neutral stances, may not fully capture the diversity of Bangla media, particularly from regional or less

prominent outlets. Additionally, the dataset relies on articles published before the models’ training cutoffs, making it challenging to ensure that models have not encountered similar content during pre-training, which could inflate performance metrics. Future work will aim to expand the dataset with more diverse sources and explore methods to verify model exposure to training data.

The task of stance detection focused on textual content, specifically article bodies, without integrating multimodal elements such as images, news-videos, or social media metadata, which often influence framing in Bangla news. We leave the exploration of multimodal stance detection to future work.

Finally, due to computational constraints, we could not perform extensive few-shot experiments, particularly proprietary ones like Claude-Sonnet-4. For consistency, we focused on models with accessible APIs or open-source weights, but future work will explore fine-tuning strategies and broader few-shot learning to improve performance, especially for neutral content detection.

References

- Gavin Abercrombie and Riza Theresa Batista-Navarro. 2020. Parlvote: A corpus for sentiment analysis of political debates. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8614–8624.
- Md Sayeed Al-Zaman and Mridha Md Shiblee Noman. 2024. Investigating political news coverage patterns in bangladesh during 2013-2022 using computational methods. *Newspaper Research Journal*, page 07395329231221515.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Pre-

- dicting the political ideology of news articles. *arXiv preprint arXiv:2010.05338*.
- Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.
- Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiusi Du, Zhe Fu, and 1 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Michael Burnham. 2025. Stance detection: a practical guide to classifying political beliefs in text. *Political Science Research and Methods*, 13(3):611–628.
- Dallas Card, Amber Boydston, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.
- Loredana Caruccio, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. 2024. Claude 2.0 large language model: Tackling a real-world classification problem with a new iterative prompt engineering approach. *Intelligent Systems with Applications*, 21:200336.
- Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International conference on natural language generation*, pages 79–88.
- Thomas Davidson, Dana Warmusley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM ’17*, pages 512–515.
- Akram Elbouanani, Evan Dufraisse, and Adrian Popescu. 2025. Analyzing political bias in llms via target-oriented sentiment classification. *arXiv preprint arXiv:2505.19776*.
- Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication*, 57(1):163–173.
- Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. *arXiv preprint arXiv:1909.02670*.
- Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a little to the left: A theory-grounded measure of political bias in large language models. *arXiv preprint arXiv:2503.16148*.
- NSA Fuhat and Juliana Abdul Wahab. 2024. Mapping media framing and politics: A bibliometric analysis across the globe. *International Journal of Academic Research in Business and Social Sciences*, 14(4):1011–1032.
- Matteo Guida, Yulia Otmakhova, Eduard Hovy, and Lea Frermann. 2025. Retain or reframe? a computational framework for the analysis of framing in news articles and reader comments. *arXiv preprint arXiv:2507.04612*.
- Fabiha Haider, Fariha Tanjim Shifat, Md Farhan Ishmam, Deeparghya Dutta Barua, Md Sakib Ul Rahman Sourove, Md Fahim, and Md Farhad Alam. 2024. Banth: A multi-label hate speech detection dataset for transliterated bangla. *arXiv preprint arXiv:2410.13281*.
- Jahidul Islam, Hisham Anand Barbhuiya, Yeasmin Akter, Sakib Rayhan, Farhana Akter, and SM Asiful Islam Saky. 2025. Conventional media credibility and current affairs news: Emergence of alternative media in hasina-era bangladesh. *International Journal of Social Science Research & Review*, 8(4):120–135.
- Sheikh Mohammad Shafiul Islam. 2016. Coverage bias of bangladesh television media: A portrayal of power and politics. *JURNAL KOMUNIKASI-MALAYSIAN JOURNAL OF COMMUNICATION*, 32(2):240–258.
- Christopher Jones. 2024. Political bias dataset: A synthetic dataset for bias detection and reduction. <https://huggingface.co/datasets/cajcodes/political-bias>.
- Parisa Jamadi Khiabani and Arkaitz Zubiaga. 2025. Cross-target stance detection: A survey of techniques, datasets, and challenges. *Expert Systems with Applications*, page 127790.
- Yungi Kim and Chanjun Park. 2024. Instatrans: An instruction-aware translation framework for non-english instruction datasets. *arXiv preprint arXiv:2410.01512*.
- Sebastian Lihammer. 2023. Semantic similarity comparison of political statements by chatgpt and political representatives.
- Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander G Hauptmann. 2006. Which side are you on? identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 109–116.

- Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.
- Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 504–514.
- Denis McQuail and Mark Deuze. 2020. Mcquail’s media and mass communication theory.
- Mohammad Mehadi Hasan, Fatema Binte Hassan, Md Al Jubair, Zobayer Ahmed, Sazzatul Yeakin, and Md Masum Billah. 2025. Bangla bert for hyperpartisan news detection: A semi-supervised and explainable ai approach. *arXiv e-prints*, pages arXiv–2507.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.
- Lynnette Hui Xian Ng, Iain J Cruickshank, and Roy Lee. 2025. Examining the influence of political bias on large language model performance in stance classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 1315–1328.
- Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.
- Tabia Tanzin Prama and Md Saiful Islam. 2025. Evaluating credibility and political bias in llms for news outlets in bangladesh. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 665–677.
- Nishat Raihan and Marcos Zampieri. 2025. Tigerllm-a family of bangla large language models. *arXiv preprint arXiv:2503.10995*.
- Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1650–1659.
- Peyman Rostami, Vahid Rahimzadeh, Ali Adibi, and Azadeh Shakery. 2025. Politisky24: Us political bluesky dataset with user stance labels. *arXiv preprint arXiv:2506.07606*.
- Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.
- Timo Spinde, Lada Rudnitskaia, Felix Hamborg, and Bela Gipp. 2021. Identification of biased terms in news articles by comparison of outlet-specific word embeddings. In *International Conference on Information*, pages 215–224. Springer.
- Arman Engin Sucu, Yixiang Zhou, Mario A Nascimento, and Tony Mullen. 2025. Exploiting contextual information to improve stance detection in informal political discourse with llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1097–1110.
- Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *Advances in Neural Information Processing Systems*, 37:34737–34774.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Gisela Vallejo, Timothy Baldwin, and Lea Frermann. 2023. Connecting the dots in news analysis: bridging the cross-disciplinary disparities in media bias and framing. *arXiv preprint arXiv:2309.08069*.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.
- Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. *arXiv preprint arXiv:2109.05325*.

	Accuracy	Govt. Critique			Neutral			Govt. Leaning			Weighted Avg.		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Nano Models													
Qwen2.5-0.5B	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	1.00	0.28	0.03	0.17	0.05
Qwen3-0.6B	0.41	0.58	0.57	0.57	0.35	0.15	0.21	0.22	0.52	0.31	0.44	0.41	0.40
Qwen3-1.7B	0.52	0.71	0.69	0.70	0.54	0.21	0.30	0.28	0.67	0.39	0.58	0.52	0.51
Compact Models													
Qwen2.5-3B	0.50	0.59	0.83	0.69	0.00	0.00	0.00	0.33	0.64	0.43	0.34	0.50	0.40
Qwen3-4B-I	0.52	0.68	0.80	0.73	0.50	0.07	0.12	0.28	0.67	0.40	0.55	0.52	0.46
Llama3.1-8B	0.55	0.58	0.88	0.70	0.50	0.10	0.16	0.45	0.55	0.49	0.53	0.55	0.47
Qwen3-4B-T	0.57	0.85	0.46	0.60	0.48	0.76	0.59	0.48	0.48	0.48	0.65	0.57	0.58
Qwen3-4B	0.58	0.73	0.68	0.71	0.56	0.42	0.48	0.39	0.67	0.49	0.61	0.58	0.59
Qwen3-8B	0.62	0.75	0.73	0.74	0.62	0.42	0.50	0.42	0.76	0.54	0.65	0.62	0.62
TigerLLM-9B	0.68	0.74	0.78	0.76	0.62	0.65	0.64	0.62	0.45	0.53	0.68	0.68	0.68
Standard Models													
Qwen2.5-14B	0.57	0.72	0.75	0.74	0.62	0.29	0.40	0.34	0.70	0.46	0.62	0.57	0.57
Qwen3-30B-I	0.62	0.70	0.89	0.78	0.77	0.24	0.36	0.41	0.70	0.52	0.68	0.62	0.59
Qwen3-14B	0.61	0.76	0.66	0.71	0.51	0.60	0.55	0.48	0.48	0.48	0.62	0.61	0.61
Qwen3-30B-T	0.61	0.85	0.56	0.68	0.51	0.64	0.57	0.50	0.73	0.59	0.67	0.61	0.62
GPT-OSS-20B	0.64	0.75	0.72	0.73	0.55	0.60	0.57	0.55	0.52	0.53	0.64	0.64	0.64
Qwen3-32B	0.68	0.82	0.68	0.75	0.55	0.72	0.63	0.70	0.58	0.63	0.71	0.68	0.68
Massive Models													
DeepSeek-V3.1	0.64	0.79	0.76	0.77	0.59	0.44	0.51	0.42	0.70	0.52	0.66	0.64	0.64
DeepSeek-R1	0.67	0.76	0.78	0.77	0.60	0.53	0.56	0.54	0.64	0.58	0.66	0.67	0.66
GPT-OSS-120B	0.67	0.76	0.78	0.77	0.59	0.51	0.55	0.55	0.67	0.60	0.66	0.67	0.66
GLM-4.5	0.69	0.82	0.72	0.76	0.60	0.69	0.65	0.59	0.61	0.60	0.70	0.69	0.69
Qwen3-235B-T	0.72	0.81	0.77	0.79	0.66	0.69	0.68	0.62	0.64	0.63	0.72	0.72	0.72
Llama3.3-70B	0.73	0.77	0.91	0.83	0.80	0.50	0.62	0.57	0.76	0.65	0.75	0.73	0.73
Qwen3-235B-I	0.74	0.78	0.86	0.82	0.71	0.62	0.67	0.66	0.64	0.65	0.74	0.74	0.74
Proprietary Models													
GPT-5-Mini	0.65	0.88	0.60	0.71	0.52	0.75	0.62	0.56	0.55	0.55	0.70	0.65	0.65
GPT-4.1-Mini	0.66	0.81	0.74	0.77	0.58	0.53	0.55	0.49	0.73	0.59	0.68	0.66	0.66
Grok-4	0.69	0.84	0.67	0.75	0.60	0.75	0.67	0.59	0.61	0.60	0.71	0.69	0.69
Gemini-2.5-Pro	0.69	0.79	0.77	0.78	0.61	0.61	0.61	0.56	0.61	0.58	0.69	0.69	0.69
Claude-Sonnet-4	0.71	0.90	0.73	0.80	0.63	0.69	0.66	0.52	0.70	0.60	0.74	0.71	0.72

Table 2: Accuracy and label-wise performance metrics of models on BanglaBias benchmark dataset. For each label (Government Critique, Neutral, Government Leaning), precision (P), recall (R), and F1-scores are reported, along with the overall weighted averages. For each block, the results are sorted based on the Weighted F1 score.

A Performance metrics of all 28 Models

Table 2 shows stance detection performance evaluation results for all 28 models (5 categories).

Article ID	Model Name	True Label	Predicted Label	Reasoning
20	Claude-Sonnet-4	Govt. Leaning	Neutral	The article provides factual reporting about Awami League’s candidate announcement for the 12th national parliament election, including specific names and constituencies. It presents information objectively without editorial commentary or bias toward or against the government.
25	Qwen3-4B-T	Neutral	Govt. Critique	The article prominently features the opposition’s (BNP) accusations against the government as ‘unlawful, unelected, and occupying the state power’ with detailed claims of human rights violations and violence. While it includes the government’s response, the narrative consistently emphasizes the opposition’s critique of the government’s actions and policies.
98	DeepSeek-V3.1	Neutral	Govt. Leaning	The article reports the government’s decision to impose a strict lockdown in a factual manner, focusing on the official announcement and the measures being taken. It presents the government as actively responding to the crisis without including significant critical perspectives or questioning the decision.
110	Qwen3-32B	Govt. Critique	Neutral	The article reports conflicting claims between protesters and the government-aligned student organization without taking a clear stance. It presents both accusations and denials objectively, focusing on factual events without evident bias.
148	Qwen3-235B-I	Neutral	Govt. Leaning	The article predominantly presents the government’s perspective, portraying BNP’s demands as unreasonable and disruptive, while highlighting state efforts to maintain peace and order. Government officials’ statements are quoted extensively to justify decisions and frame BNP’s actions as potentially inciting chaos.

Table 3: Selected cases of misclassified articles in our Bangla political stance detection study. Each row lists the article ID, the model used, the true label, the predicted label, and the reasoning provided by the model.

B Examples of Wrongly Predicted Label and Model-Reasoning

Table 3 shows five error examples along with models’ reasoning.

News Source	Article Count	News Source	Article Count
BBC Bangla	37	Doinik Bangla	1
Prothom Alo	20	Bagerhatinfo	1
Jugantor	17	USEmbassy	1
DW	11	Dainik Shiksha	1
Bangla Tribune	11	Bonik Barta	1
The Daily Star	10	Bangladesh Jamate Islam	1
Samakal	7	Kaler Kontho	1
Bdnews24	7	Bdnew24	1
Dhaka Post	7	News Bangla	1
Bangla News24	7	Bangla 24 live newspaper	1
Jago News	6	Dhaka Times	1
Somoy news	5	SattAcademy	1
Dhaka Times	5	voabangla	1
Daily Ittefaq	4	Daily Janakantho	1
Daily Inqilab	4	Ajker Potrika	1
Dhaka Tribune	2	Desh Rupantor	1
Somoyer Alo	2	Daily Campus	1
Channel online	2	Khulna Gazet	1
The Daily Vorer Pata	1	Cvoice24	1
Ajkaler khobor	1	Dainik Sylhet	1
BanglaTribune	1	Dainik amader bangla	1
smsaif	1	Chalaman neywork	1
somewhereinblog	1	The doctors dialogue	1
Timetouch news	1	Kaler kontho	1
Banik Barta	1	somewhere in blog	1
DBC news	1	rajibkhaja	1
MuktiBarta	1	Green Watch BD	1

Table 4: Distribution of articles across news sources, showing the number of collected articles from each outlet.

C News Sources and Respective Article Count

Table 4 includes all the news sources and the number of articles sourced from respective corpora, blog or site.

Event Name	Event Name
1 তাজরীন ফ্যাশন অগ্নিকাণ্ড	24 কোভিড১৯ গোটা দেশে শুরু
2 ১৯৯৬ একতরফা নির্বাচন	25 গার্মেন্টস খোলা ও শ্রমিক মজুরি
3 সোনালী ব্যাংক হলমার্ক ঋণ কেলেঙ্কারি	26 ধর্ষণবিরোধী আন্দোলন ও মৃত্যুদণ্ড আইন
4 শাহবাগ আন্দোলন	27 মোদির বাংলাদেশ সফর বিরোধিতা
5 হেফাজতে ইসলামের লংমার্চ ও ঢাকা অবরোধ	28 লেখক মুশতাক আহমেদ কারাগারে মৃত্যু
6 রানা প্লাজা ধস	29 পদ্মা সেতু সংযোগ সম্পন্ন
7 ২০১৪ সালের সংসদ নির্বাচন	30 সাংবাদিক রোজিনা ইসলাম মামলা
8 আইনশৃঙ্খলা বাহিনী কর্তৃক বিচারবহির্ভূত 'ক্রশফায়ার'	31 জ্বালানি মূল্য বৃদ্ধি
9 বিএনপি পরিবহন অবরোধ	32 রামপাল ইউনিট-১ চালু
10 রামপাল কয়লা বিদ্যুৎ কেন্দ্র চুক্তি	33 যুক্তরাষ্ট্রের নিষেধাজ্ঞা ও র‍্যাব বিতর্ক
11 টাঙ্গাইল ইউনিয়ন পৌর সংঘর্ষ	34 বিএনপি ঢাকা মহাসমাবেশ ও গ্রেপ্তার
12 হলি আর্টিজান হামলা	35 রাজশাহী বিশ্ববিদ্যালয় ছাত্রদল সংঘর্ষ
13 বেসরকারি বিশ্ববিদ্যালয়ে ভ্যাট আন্দোলন	36 বাজেটে ভ্যাট বাড়ানো
14 রামপাল পরিবেশ আন্দোলন বৃদ্ধি	37 রামপাল ইউনিট-২ সম্পন্ন
15 রোহিঙ্গা শরণার্থী আসা	38 ডিজিটাল আইন সংশোধনী বিল
16 ডিজিটাল নিরাপত্তা আইন পাস	39 বিএনপি ও আওয়ামী লীগ 'গ্র্যান্ড র্যালি'
17 কোটা সংস্কার আন্দোলন	40 দ্বাদশ সংসদ নির্বাচন (বয়কটসহ)
18 ছাত্রদের নিরাপদ সড়ক আন্দোলন	41 পেনশন সংস্কার বিল
19 ডিএসএ (ডিজিটাল আইনের) কার্যকর	42 সুপ্রিম কোর্ট ৫৬% কোটা পুনর্বহাল
20 ২০১৮ সালের সংসদ নির্বাচন	43 কোটা আন্দোলন: পুলিশ গুলিতে ২০০+ নিহত
21 'ক্যাসিনো রাজ্য' অভিযান	44 ইন্টারনেট বন্ধ ও কারফিউ
22 নুসরাত হত্যাকাণ্ড	45 সরকারি চাকরিতে কোটা কমিয়ে ৫%
23 আবরার ফাহাদ হত্যা	46 দেশব্যাপী 'মার্চ অন ঢাকা' ডাকা

Figure 5: List of 46 politically debatable events (spanning diverse news coverage) included in our benchmark dataset to capture multiple perspectives.

D Politically Debated Events Covered by BanglaBias

Fig. 5 includes the list of all the events covered in BanglaBias.

Short Alias	Model ID	Category
Qwen2.5-0.5B	Qwen_Qwen2.5-0.5B-Instruct	Nano
Qwen3-0.6B	Qwen_Qwen3-0.6B	Nano
Qwen3-1.7B	Qwen_Qwen3-1.7B	Nano
Qwen2.5-3B	Qwen_Qwen2.5-3B-Instruct	Compact
Qwen3-4B	Qwen_Qwen3-4B	Compact
Qwen3-4B-I	Qwen_Qwen3-4B-Instruct-2507	Compact
Qwen3-4B-T	Qwen_Qwen3-4B-Thinking-2507	Compact
Qwen3-8B	Qwen_Qwen3-8B	Compact
Llama3.1-8B	meta-llama_Llama-3.1-8B-Instruct	Compact
TigerLLM-9B	md-nishat-008_TigerLLM-9B-it	Compact
Qwen2.5-14B	Qwen_Qwen2.5-14B-Instruct	Standard
Qwen3-14B	Qwen_Qwen3-14B	Standard
Qwen3-30B-I	Qwen_Qwen3-30B-A3B-Instruct-2507	Standard
Qwen3-30B-T	Qwen_Qwen3-30B-A3B-Thinking-2507	Standard
Qwen3-32B	Qwen_Qwen3-32B	Standard
GPT-OSS-20B	openai_gpt-oss-20b	Standard
Llama3.3-70B	meta-llama_Llama-3.3-70B-Instruct	Massive
GPT-OSS-120B	openai_gpt-oss-120b	Massive
Qwen3-235B-I	qwen_qwen3-235b-a22b-Instruct-2507	Massive
Qwen3-235B-T	qwen_qwen3-235b-a22b-Thinking-2507	Massive
GLM-4.5	z-ai_glm-4.5	Massive
DeepSeek-V3.1	deepseek_deepseek-v3.1-terminus	Massive
DeepSeek-R1	deepseek_deepseek-r1	Massive
Grok-4	x-ai_grok-4-fast	Proprietary
Claude-Sonnet-4	anthropic:claude-sonnet-4	Proprietary
Gemini-2.5-Pro	google:gemini-2.5-pro	Proprietary
GPT-4.1-Mini	openai:gpt-4.1-mini	Proprietary
GPT-5-Mini	openai:gpt-5-mini	Proprietary

Table 5: Mapping between short aliases and full model id, along with their assigned model category (Nano, Compact, Standard, Massive, or Proprietary).

E Model ID Mapping

Table 5 contains the detailed model id and categories.

F Human Annotator Details

Our annotators were graduate students from three different academic backgrounds. Each had a strong understanding of the Bangladesh's socio-political context and actively followed national media coverage. We deliberately selected annotators from diverse majors to foster inclusive knowledge exchange and ensure a multi-perspective approach to the task. None of them had any professional or institutional ties to media outlets. The annotators were fully informed about the purpose of the project and how their annotations would be used.

G Evaluation Prompt

Zero-shot evaluation prompt

You are a political bias classifier.

Task:

Read the following article and return ONLY a valid JSON object with two fields:

```
{{
  "label": one of ["Govt leaning",
    "Neutral", "Govt critique"],
  "reason": "1-3 sentences explanation"
}}
```

Rules:

- Output ONLY JSON.
- Do not include markdown, code fences, or explanations.

Article:

```
{{article_text}}
```

Few-shot evaluation prompt

You are a political bias classifier.

Task:

Read the following article and return only a valid JSON object with two fields:

```
{{{{
  "label": one of ["Govt leaning",
    "Neutral", "Govt critique"],
  "reason": "1-3 sentences explanation"
}}}}
```

Rules:

Output ONLY JSON (no markdown, no extra text).

Base your decision strictly on tone, stance, and framing not just factual accuracy.

Examples:

Example 1:

Article:

...

```
...
Output:
{{"label": "Neutral", "reason": "..."}}
Example 2:
Article:
...
...
Output:
{{"label": "Govt leaning", "reason": "..."}}
Example 3:
Article:
...
...
Output:
{{"label": "Govt
critique", "reason": "..."}}

Now classify the following article:
{article_text}
```

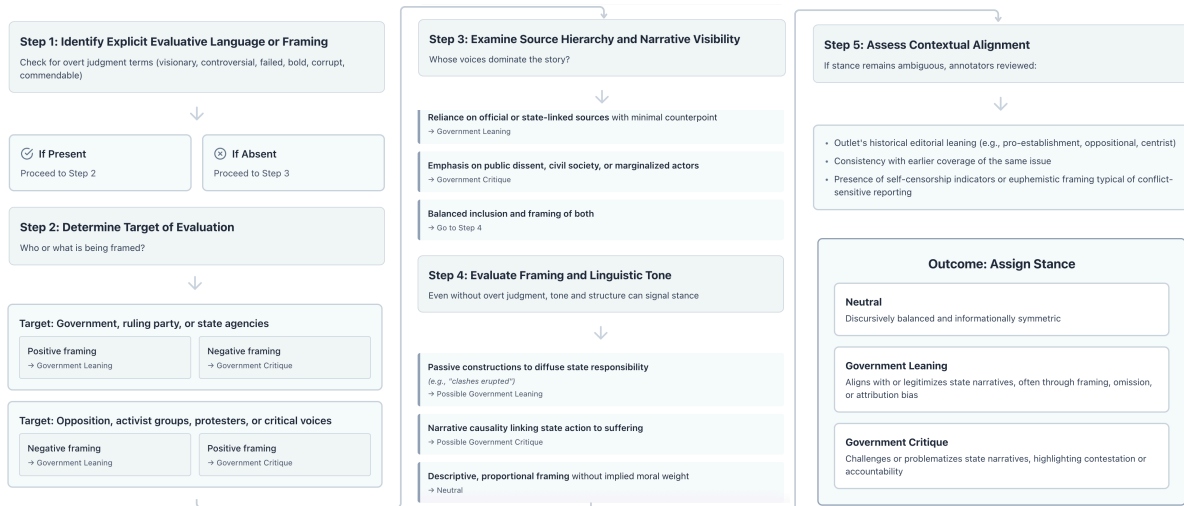


Figure 6: Annotation Decision Flowchart for Political Stance Classification in the Bangladeshi Context. The decision flowchart operationalizes stance annotation within a politically polarized media system, where neutrality is a negotiated discursive position rather than a natural default.

H Annotation Decision Flowchart

A high-level annotation decision flowchart is provided in the Fig. 6.

Model	Govt. Critique			Neutral			Govt. Leaning			Weighted Avg		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Qwen3-1.7B (Few-Shot)	0.66	0.78	0.71	0.56	0.60	0.58	0.55	0.18	0.27	0.60	0.62	0.59
Qwen3-1.7B (Zero-Shot)	0.71	0.69	0.70	0.54	0.21	0.30	0.28	0.67	0.39	0.58	0.52	0.51
TigerLLM-9B (Few-Shot)	0.86	0.58	0.69	0.51	0.82	0.63	0.57	0.36	0.44	0.69	0.63	0.63
TigerLLM-9B (Zero-Shot)	0.74	0.78	0.76	0.62	0.65	0.64	0.62	0.45	0.53	0.68	0.68	0.68
Qwen3-32B (Few-Shot)	0.84	0.45	0.59	0.45	0.76	0.57	0.54	0.45	0.49	0.65	0.57	0.57
Qwen3-32B (Zero-Shot)	0.82	0.68	0.75	0.55	0.72	0.63	0.70	0.58	0.63	0.71	0.68	0.68
Llama3.3-70B (Few-Shot)	0.86	0.68	0.76	0.59	0.57	0.58	0.49	0.82	0.61	0.70	0.67	0.67
Llama3.3-70B (Zero-Shot)	0.77	0.91	0.83	0.80	0.50	0.62	0.57	0.76	0.65	0.75	0.73	0.73

Table 6: Few-Shot vs Zero-Shot Performance Comparison

I Zero-shot vs Few-shot

Zero-shot vs few-shot prompting is provided in Table 6.