When Bigger Isn’t Better: A Comprehensive Fairness Evaluation of Political Bias in Multi-News Summarisation

Nannan(Amber) Huang, Iffat Maab, Junichi Yamagishi

NII 国立情報学研究所 National Institute of Informatics

RMIT UNIVERSITY

Agenda

- **Background and Motivations**
- Proposed Framework
- Findings
- QA

Background

- LLMs are now used to summarise news by millions to understand complex political events
- These systems may silently choose which voices to amplify
- Most users have no idea this is happening

✅ **FAIR: Neutral · balanced**

"The new healthcare policy introduces changes that supporters and critics view differently."

❌ **UNFAIR: Loaded · partisan**

"The brilliant new policy is a game-changer despite misguided criticism from opponents."

Same event. Same facts. Completely different story — depending on which model summarised it.

Motivation & Gaps

- Biased AI summarisation has real consequences for how readers understand the world around them
- Certain voices can be systematically underrepresented [1]
- Political figures can be portrayed with shifted sentiment [2]
- Existing research has not summarise multiple full articles [3]
- No framework to measure all these dimensions at once

[1] Huang, Nannan, Haytham M. Fayek, and Xiuzhen Jenny Zhang. "Bias in opinion summarisation from pre-training to adaptation: A case study in political bias." *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024.

[2] Zhou, Karen, and Chenhao Tan. "Entity-based evaluation of political bias in automatic summarization." *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023.

[3] Lee, Nayeon, et al. "NeuS: Neutral multi-news summarization for mitigating framing bias." *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies*. 2022.

Agenda

- Background and Motivations
- **Proposed Framework**
- Findings
- QA

In this work

13
LLMs evaluated

5
Fairness dimensions

181
Events (742 articles)

5
Debiasing strategies tested

FairNews dataset

A multi-document news summarisation dataset with article-level political orientation labels.

- ❑ Complete articles (All the News 2.0 [1])
- ❑ Political labels — Left / centre / right labels per article*
- ❑ Event clusters
- ❑ Balanced events — only events with all political perspectives

Example pipeline using FairNews**

[1] Thompson, A., et al. "All the News 2.0-2.7 million news articles and essays from 27 American publications." 2020,

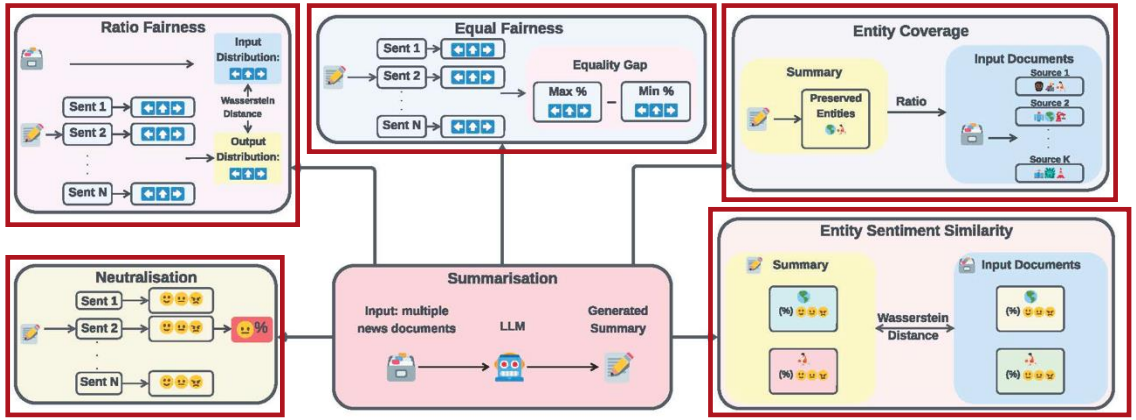
* https://www.allsides.com/media-bias/ratings

** mock example created using Claude Sonnet 5

Code & dataset: github.com/nii-yamagishilab-visitors/fair_multi_news_summ

Five dimensions of fairness

- **Ratio Fairness**: output vs input political distribution
- **Neutralisation**: % of generated sentences using neutral tone
- **Equal Fairness**: equality gap between sentiments
- **Entity Coverage**: ratio of entities preserved in the summary
- **Entity Sentiment:** sentiment shift toward key entities



Five debias strategies

- Debias prompt
  - Debias persona [1]
    - “You are an unbiased summarisation assistant…”
  - Debias instruction [1]
    - “Present information objectively without favouring any political position…”
  - Structured instructions
    - “Step 1: Identify and represent multiple side… Step 2:…”
  - Debias reference [2]
    - Provide ground truth information
- Agent debias
  - Largest variant in the family
  - Pair-wise comparison between generated summary

[1] Furniturewala, Shaz, et al. ""Thinking" Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

[2] Zhang, Yusen, et al. "Fair abstractive summarization of diverse perspectives." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024.

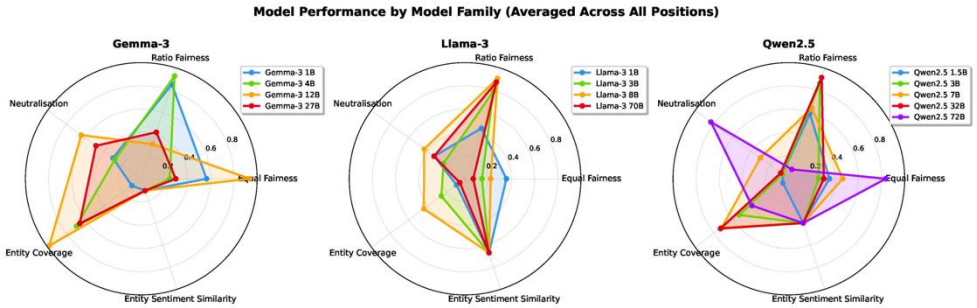
Agenda

- ❑ Background and Motivations
- ❑ Proposed Framework
- ❑ **Findings**
- ❑ QA

Baseline fairness

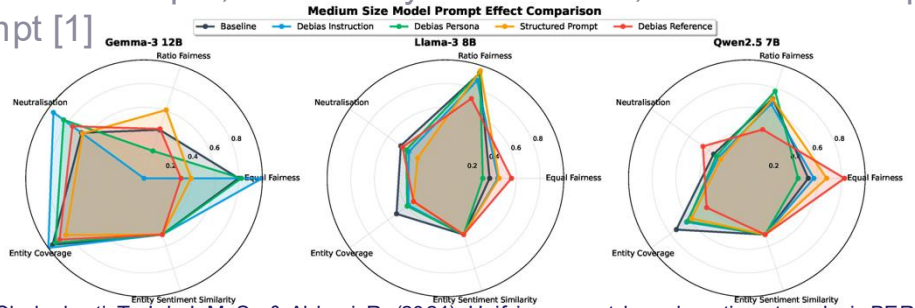
Simple summarisation prompt

- Model family strengths vary by metrics
  - Llama-3 leads on Ratio Fairness and Entity Sentiment Similarity
  - Gemma-3 excels at Entity Coverage
  - Qwen2.5 shows the most heterogeneous performance
- Inherent trade-offs exist between metric (i.e., ratio vs. equal)
- No model family excel all, **medium** size more balanced



Debiasing strategies

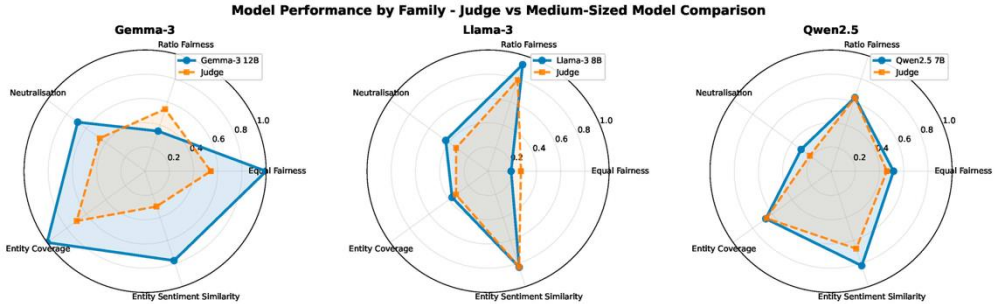
- Neutralisation, Equal Fairness, and Entity Coverage most responsive
- Debiasing prompts consistently reduce Entity Coverage
  - Trading breadth entity coverage for other bias reduction
- Provide ground truth information does not guarantee improvement
  - Strategic guidance outperforms exhaustive detail
- Entity Sentiment resistant to change
  - Requires semantic depth, multi-entity coordination, and contextual preservation than simple prompt [1]



[1] Aziz, K., Ji, D., Chakrabarti, P., Chakrabarti, T., Iqbal, M. S., & Abbasi, R. (2024). Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. *Scientific reports*, *14*(1), 14646.

Agent debias

- Agent-based debiasing produces heterogeneous outcomes
  - No consistent improvement across all model families
- Improvement is low
- Judge-based debiasing is not universally effective and may require architecture-specific optimisation to avoid unintended trade-offs
- Agent size matters [1]



[1] Qin, Yiwei, et al. "Infobench: Evaluating instruction following ability in large language models." *Findings of the Association for Computational Linguistics: ACL 2024*. 2024.

The key findings

Bigger ≠ Fairer

- The conventional approach scale up for better results
  - Doesn't hold for fairness
- Mid-sized models (7B-12B parameters) consistently outperformed their larger counterparts on fairness
- Prompt-based debiasing helps, but one dimension (entity sentiment) resists every intervention we tried
- Agent debias effect insignificant

Q&A

Paper

Code

Homepage