Hidden bias in the pursuit of truth: A longitudinal analysis of political neutrality in independent fact-checking

Sahajpreet Singh* *National University of Singapore*
sahajpreet.singh@u.nus.edu
Sarah Masud* *University of Copenhagen* sarah.masud@di.ku.dk
Tanmoy Chakraborty *Indian Institute of Technology Delhi* tanchak@iitd.ac.in

Abstract

Independent fact-checking organisations have emerged as the guardians against fake news. However, these organisations might deviate from political neutrality by being selective in what false news they debunk and how the debunked information is presented. At the intersection of AI for social science and humanities, this work explores how journalistic frameworks and large language models (LLMs) can be employed to detect political biases in fact-checking organisations and make the public aware of the limitations of such a setup.

Prompting GPT-3.5, with journalistic frameworks of 5W1H, we establish a longitudinal measure (2018-2023) for political neutrality that looks beyond the left-right spectrum. Specified on a range of -1 to 1 (with zero being absolute neutrality), we establish the extent of negative portrayal of political entities that marks a difference in the readers' perception in six independent fact-checking organisations across the USA and India. Here, we observe an average score of -0.17 and -0.24 in the USA and India, respectively. The findings indicate how seemingly objective fact-checking can still carry distorted political views, indirectly and subtly impacting consumers' perception of the debunked news.

Keywords: Fact-checking, political neutrality, and AI in social computing.

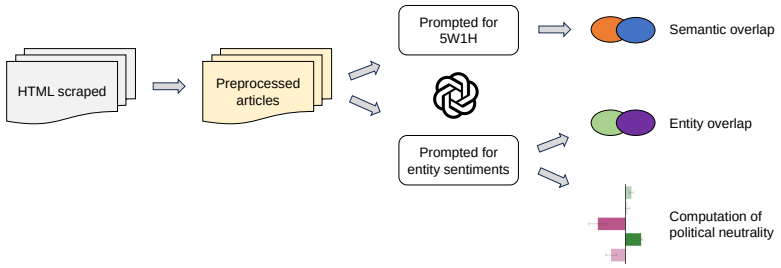

Figure 1: **Flowchart of the proposed setup.** Scraping raw text from fact-checking organisations, and preprocessing is followed by GPT-based prompting to obtain 5W1H and entity-based sentiment mapping. The prompt outputs are employed to compute political neutrality and the entity and semantic overlap.

AI in Science Conference 2025
EU2025DK

1 Overview

In this work, we define independent fact-checking organisations as online platforms that are neither affiliated with nor funded by any conventional news media or government. The factual/objective nature of the work done by independent fact-checking organisations obfuscates the need to examine biases and political leanings among these organisations. Further, the limitations of left-right binarisation of conventional media platforms Kim et al. [2022], Guess et al. [2021], Eberl et al. [2017] call for a broader examination of political leaning Flamino et al. [2023] globally. Inspired by these limitations, we attempt to examine how the biases manifest in independent fact-checking organisations. Our research builds on the assistance of LLMs for generating pseudo labels and acting as sources of information summarisation Sundriyal et al. [2023], Shah et al. [2025]. While journalistic framing for fact-verification/classification has been explored Rani et al. [2023], we uniquely combine journalistic question-answering with sentiment subjectivity in a free-text format to evaluate neutrality.

Hypothesis. *We hypothesise that the disparity in neutrality among organisations is an amalgamation of topical and narrative accentuations Lee et al. [2023], Stevenson et al. [1973] in the presentation of fact-check news articles.*

Methodology. We perform a longitudinal investigation from circa 2018 to 2023. To further validate the generalizability of our technique beyond geography and the political spectrum, we employ India and the USA as representative countries. The process begins with the curation of $\sim 25k$ and $\sim 10k$ fact-checking articles from six prominent independent fact-checking organisations[1]. Here, we employ Check Your Fact, PolitiFact, Snopes for the USA and Alt News, Boom, and OpIndia for India. Employing journalistic practices of 5W1H (What, Why, Where, When, Who, How) Bleyer [1913] in conjunction with prompting capabilities of GPT-3.5-based large language model (LLM) Kojima et al. [2022], we quantify the extent of political neutrality exhibited by these organisations. We uniquely express political neutrality on a continuous scale from -1 to 1, aggregated over political entities mentioned in each article. Here, we prompt GPT-3.5 to extract the political entities/agencies (political parties and leaders) mentioned in news articles and summarise their neutrality using perception tags (negative, neutral, or positive). This process is akin to aspect-based sentiment analysis, wherein the same article can convey varying portrayals of multiple political entities. The approach is summarised in Figure 1.

Limitations. It should be noted that the proposed methodology is not infallible. Hence, our proof of concept accounts for both statistical uncertainty measurement of polarity scores and human evaluation of the GPT-prompted results. As our work focuses on the semantic and retrieval capabilities of the LLMs and does not prompt the LLM to generate new opinions, we can reduce the impact of hallucinations of LLMs Wright et al. [2024].

[1]The research is not funded by any of the fact-checking organisations or political parties examined in this study.

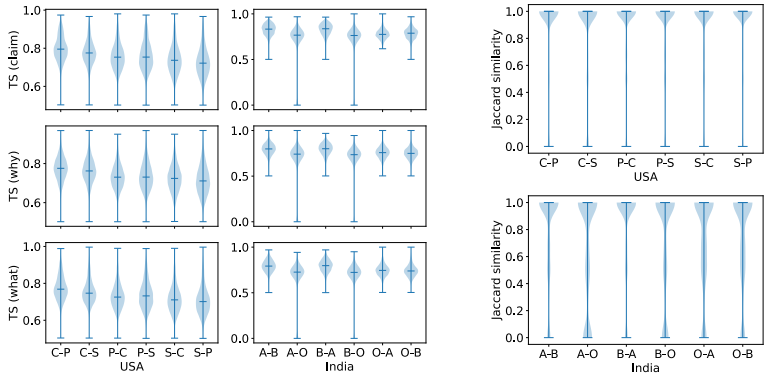

Figure 2: The data from six independent fact-checking organisations curated over five years (2018-2023) – PolitiFact (P), Snopes (S), and Check Your Fact (C) from the USA, and Alt News (A), OpIndia (O), and Boom (B) from India. Within a timeframe of ± 15 days, for news articles of organisation X with organisation Y, we report: (a) Inter-organisation (X-Y) maximum topical similarity (TS) and similarity threshold 0.75. The max TS values are recorded for the "Claim", "What," and "Why" tags. (b) Inter-organisation (X-Y) Jaccard similarity over the $top_k = 100$ political entities. Note: We record the similarity only among organisations within a geography.

2 Observations

Notably, from Figures 2 and 3, the skewness in topical versus entity similarity shows that while the same entities are discussed, "what" about these entities covered and "how" they are represented varies. The skewness and subsequent representation are indicators of media bias. *By not eulogising and criticising political entities equally, fact-checking organisations risk the loss of neutral behaviour.* They exhibit a form of media bias stemming from inequity of coverage rather than a departure from truth Hamborg [2023]. This portrayal of favouritism feeds the vicious cycle of jaundiced perception among the readers, with direct implications for the unassuming readers who depend on such organisations for objectivity.

A more coherent left-right political ideology of the USA vs India comes into play via compelling dynamics; the person-party alignment in the overall image is better aligned in the USA than in India. It is most evident in terms of the head of state. *For both the heads of state (Modi and Biden as of our 2023 cutoff), PS is mostly* ≥ 0*, which may not always align with the portrayal of their parent political party, even with the same news organisation.* This observation validates the use of neutrality-based image portrayal, looking beyond the left-right political spectrum Puthillam et al. [2021]. Despite differences in political landscapes, our LLM-enhanced setup allows us to examine political biases beyond the myopic view of left vs right. The similarity in patterns of political neutrality or lack thereof in fact-checking

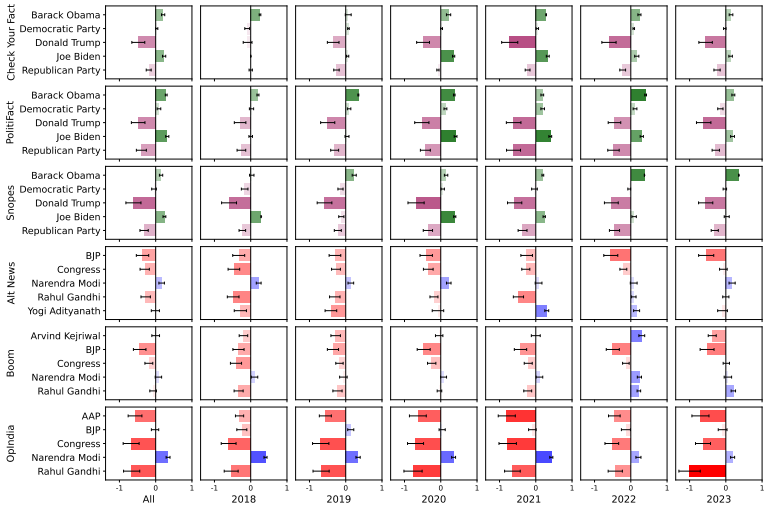

Figure 3: The extent of neutrality ($-1 \leq PS \leq 1$) for the $top_k = 5$ entities per fact-checking organisation. PolitiFact, Snopes, and Check Your Fact are in the USA. Alt News, OpIndia, and Boom are based in India. PS establishes "how" *the coverage of the political entities in the fake and debunked news leads to positive, negative, or neutral imagery impacts the reader's perception.* A higher neutrality is observed if $PS \approx 0$. Meanwhile, a score closer to -1 (1) highlights a more pessimistic/critical (positive/promoting) tone in portraying the entities. Note: The log error bars account for the uncertainty in PS.

of both the USA and India further corroborates the need for the proposed setup.

3 Impact on Consumers of News

From the perspective of readers of the debunked news, three critical observations stand out. We hope our observations allow users to be cognizant of political biases in independent fact-checking organisations:

- Given that within a country, all organisations focus on the same set of political entities, on the surface, a negligible predisposition will be visible to the reader.
- However small, the values in terms of neutrality scores (-0.28 to -0.10 and -0.27 to -0.19) indicate that the difference in the narrative is subtle and yet has the potential to mislead unintended readers of the news.
- The overall perception meted out to entities by an organisation remains similar over the five years. Thus, no surprising political shifts will appear for a reader who is a long-time follower of these organisations.

4 Error Estimation and Output Validation

4.1 Human evaluation of LLM generated outputs

We employ the help of two expert annotators who volunteer to evaluate our setup. The annotators know American and Indian politics and understand how fact-checking works. The annotators are one male and one female, aged 20-30, familiar with computational social science and prompt engineering in LLMs. The annotators are randomly provided fact-checked news items from the six organisations (with references to the organisations anonymised to reduce bias) along with the outputs obtained from prompts. The annotators independently assess 20 samples and evaluate the following:

- **Relevancy score for retrieved What, Why, and Claim:** We ask annotators to score the retrieved sentences for relevance on a Likert scale between 0 and 2 with 0 being irrelevant, 1 being somewhat relevant, and 2 being highly relevant. For evaluator 1, the average relevancy scores were 1.75 for What, 1.67 for Why, and 1.87 for Claim, while for evaluator 2, the scores were 1.67 for What, 1.33 for Why, and 1.61 for Claim.
- **Recall of extracted entities and precision of generated perception tags:** We ask human annotators to manually identify the major political entities in the articles and assign each one a perception tag of positive, negative, or neutral. Based on this, we find the average recall of relevant political entities to be 93.3% for one evaluator and 87.5% for another. The average precision (same for both evaluators) of the annotated tags is 90.2%. Specifically, the precision for positive and neutral classes is perfect, while the precision for the negative class is 70.6%. We use this precision to find the uncertainty using maximum log error for the polarity score.

4.2 Significance and uncertainty measurement

Confidence interval for topical similarity. To assess the significance of the topical similarity between two organisations, X and Y, we employ the bootstrap method to calculate the 95% confidence interval for the median of the topical similarity (TS). We generate 10,000 bootstrap samples, each consisting of 20% of the total number of articles from organisation X.

Maximum log error in polarity scoring. We add the maximum log error bar to quantify the maximum propagated error resulting from these prediction inaccuracies. Here, for a single entity, N_p and N_n denote the total number of positive and negative mentions, respectively, out of the total occurrence for an entity N_t, we have $PS = \frac{N_p - N_n}{N_t}$. The error difference can be simplified as Equation 1. Here, Pr represent Precision, and this can be understood with the intuition that if we take $1 - Pr$ as an error in a single instance, the total error will be a multiple of $N_{p/n}$.

$$\Delta PS = \frac{N_p \times (1 - Pr_p) + N_n \times (1 - Pr_n)}{N_t} \tag{1}$$

References

W. G. Bleyer. *Newspaper writing and editing*. Houghton Mifflin, Boston, USA, 1913.

J.-M. Eberl, H. G. Boomgaarden, and M. Wagner. One bias fits all? three types of media bias and their effects on party preferences. *Communication Research*, 44(8):1125–1148, 2017. doi: 10.1177/0093650215614364. URL https://doi.org/10.1177/0093650215614364.

J. Flamino, A. Galeazzi, S. Feldman, M. W. Macy, B. Cross, Z. Zhou, M. Serafino, A. Bovet, H. A. Makse, and B. K. Szymanski. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. *Nature Human Behaviour*, 7(6):904–916, Jun 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01550-8. URL https://doi.org/10.1038/s41562-023-01550-8.

A. M. Guess, P. Barberá, S. Munzert, and J. Yang. The consequences of online partisan media. *Proceedings of the National Academy of Sciences*, 118(14):e2013464118, 2021. doi: 10.1073/pnas.2013464118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2013464118.

F. Hamborg. *Media Bias Analysis*, pages 11–53. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-17693-7.

E. Kim, Y. Lelkes, and J. McCrain. Measuring dynamic media bias. *Proceedings of the National Academy of Sciences*, 119 (32):e2202197119, 2022. doi: 10.1073/pnas.2202197119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2202197119.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

S. Lee, A. Xiong, H. Seo, and D. Lee. "fact-checking" fact checkers: A data-driven approach. *Harvard Kennedy School Misinformation Review*, 2023.

A. Puthillam, H. Kapoor, and S. Karandikar. Beyond left and right: A scale to measure political ideology in india, Feb 2021. URL osf.io/preprints/psyarxiv/fg387.

A. Rani, S. T. I. Tonmoy, D. Dalal, S. Gautam, M. Chakraborty, A. Chadha, A. Sheth, and A. Das. FACTIFY-5WQA: 5W aspect-based fact verification through question answering. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *ACL*, pages 10421–10440, Toronto, Canada, July 2023. ACL. doi: 10.18653/v1/2023.acl-long.581. URL https://aclanthology.org/2023.acl-long.581.

B. S. Shah, D. S. Shah, and V. Attar. Decoding news bias: Multi bias detection in news articles. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 97–104, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400717383. doi: 10.1145/3711542.3711601. URL https://doi.org/10.1145/3711542.3711601.

R. L. Stevenson, R. A. Eisinger, B. M. Feinberg, and A. B. Kotok. Untwisting the news twisters: A replication of efron's study. *Journalism Quarterly*, 50(2):211–219, 1973.

M. Sundriyal, T. Chakraborty, and P. Nakov. From chaos to clarity: Claim normalization to empower fact-checking. In *Findings of EMNLP*, pages 6594–6609, 2023.

D. Wright, A. Arora, N. Borenstein, S. Yadav, S. Belongie, and I. Augenstein. LLM tropes: Revealing fine-grained values and opinions in large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.995. URL https://aclanthology.org/2024.findings-emnlp.995/.