Findings of the Shared Task Multilingual Bias and Propaganda Annotation in Political Discourse

Shunmuga Priya Muthusamy Chinnan[1], Bharathi Raja Chakravarthi[1], Meghann L. Drury-Grogan[3], Senthil Kumar B[4], Saranya Rajiakodi[5], Angel Deborah Suseelan[6], Jason Joachim Carvalho[1]

[1] School of Computer Science, University of Galway, Ireland
[3] Department of Enterprise and Technology, Atlantic Technological University, Ireland
[4] Department of Information Technology, Velammal Institute of Technology, India
[5] Department of Computer Science, Central University of Tamil Nadu, India
[6] Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, India

Abstract

The Multilingual Bias and Propaganda Annotation task focuses on annotating biased and propagandist content in political discourse across English and Tamil. This paper presents the findings of the shared task on bias and propaganda annotation task. This task involves two sub tasks, one in English and another in Tamil, both of which are annotation task where a text comment is to be labeled. With a particular emphasis on polarizing policy debates such as the US Gender Policy and India's Three Language Policy, this shared task invites participants to build annotation systems capable of labeling textual bias and propaganda. The dataset was curated by collecting comments from YouTube videos. Our curated dataset consists of 13,010 English sentences on US Gender Policy, Russia-Ukraine War and 5,880 Tamil sentences on Three Language Policy. Participants were instructed to annotate following the guidelines at sentence level with the bias labels that are fine-grained, domain specific and 4 propaganda labels. Participants were encouraged to leverage existing tools or develop novel approaches to perform fine-grained annotations that capture the complex socio-political nuances present in the data.

1 Introduction

Social media platforms have become important medium for communication, enabling the widespread exchange and access to information from diverse sources (Datta et al., 2021). However, this open ecosystem is increasingly filled with harmful content, including various forms of misinformation such as propaganda, conspiracy theories, and biased narratives (Zubiaga et al., 2016). The rapid scale and sophistication of such content demand solutions beyond manual fact checking (Nakov and Da San Martino, 2020). Consequently, developing automated methods to detect and mitigate biased and propagandist content has become an urgent research priority (Zaghouani et al., 2024; Aksenov et al., 2021).

The term bias is defined as "Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is inaccurate, closed-minded, prejudicial, or unfair" (Steinbock, 1978). Biases can be innate or learned (Welsh and Begg, 2016). Propaganda can take many forms, including political speeches, advertisements, news reports, and social media posts (Guess and Lyons, 2020). Its goal is usually to influence people's attitudes and behaviors, either by promoting a particular ideology or by persuading them to take a specific action (Berinsky, 2017; Casavantes et al., 2024).

Hence our task [1], [2], addresses the critical need for analyzing bias and propaganda in multilingual political discourse. Developing annotation guidelines for complex data is a challenging task. In our tasks, we have identified the ideological bias related to support or against the government decisions on transgender rights, three language policies. We consider this as biases because it highly judgmental on sensitivity concerns. Annotating such bias is crucial for understanding how regional and linguistic identities are influenced in political discourse (Aksenov et al., 2021). The purpose of this annotation is to examine how political narratives shapes the public opinion by favoring or attacking specific policies and identities. To overcome these limitations, annotation efforts should incorporate diverse human perspectives: involving annotators from multiple cultural and linguistic backgrounds has been shown to reduce bias and capture nuanced interpretation. Addressing this challenge requires incorporating diverse perspectives and multi cultural insights during annotation, which can significantly enhance the robustness and fairness of NLP systems.

[1] https://sites.google.com/view/lt-edi-2025/tasks?authuser=0

[2] https://codalab.lisn.upsaclay.fr/competitions/22054

2 Related Works

Understanding and detecting bias in political discourse has become main concerns in computational social science (Heppell et al., 2023). Earlier research have explored the linguistic features of biased content and propaganda tactics in news articles, speeches and online comments (Lim et al., 2020; Allcott and Gentzkow, 2017). (Rashkin et al., 2017) analyzed linguistic pattern across different types of biased text, including fake news and political fact-checks. The work identified the subtle forms of bias through lexical and syntactical structures. (Da San Martino et al., 2019) offered a fine-grained taxonomy for identifying propaganda techniques in news articles. This work emphasized on detecting 18 specific propaganda techniques in news articles, such as appeal to fear, flag-waving, and loaded language. (Baly et al., 2020) presented study on predicting the political ideology of news articles. With a comprehensive dataset of with 34,737 news articles yielded the model's robustness. The authors suggested novel modeling approaches, such as a specially modified triplet loss function and adversarial media adaptation to deal with propaganda tactics in cultural contexts.

Multilingual bias detection presented by (Maity et al., 2024) created two large-scale datasets, mWikiBias and mWNC in eight Indian languages. The authors propose techniques for the detection of neutrality bias in politically and socially sensitive articles through models such as mDeBERTa and mT5. (Chavan and Kane, 2022) proposed a method for multi label propaganda detection using LLM. The WANLP 2022 shared task, which called for recognizing several propaganda strategies in a single text, inspired the development of their system. They achieved a micro-F1 score of 59.73% by using an ensemble of five models to handle the complexity of detecting 21 different propaganda techniques. (Zaghouani et al., 2024) conducted FIGNEWS shared task as a component of the ArabicNLP 2024 conference, which was held concurrently with ACL 2024. This work used the early stages of the Israel War on Gaza as a case study to examine bias and propaganda annotation in multilingual news posts. Their findings highlights the importance of clear guidelines and collaborative efforts in advancing NLP research on sensitive opinion analysis tasks.

3 Task Description

The shared task, addresses the crucial need for analyzing bias and propaganda in multilingual political discourse. This task aligns with the NLP community growing efforts to create datasets and guidelines for complex opinion analysis through collaborative shared tasks. There are two tasks in this shared task

- Task 1: Bias and Propaganda Annotation in English
  - Sub Task 1.1: US gender policy dataset The goal is to focuses on annotating contents related to Trump's US gender policy against transgender individuals. The task is to annotate based on the bias and propaganda guidelines in English texts that discuss this policy. There are totally 6 bias labels and 4 propaganda labels.
  - Sub task 1.2: Russia-Ukraine dataset Annotate the content of YouTube comments related to the Ukraine-Russia war in English. The task involves categorizing the comments based on bias and propaganda, following established guidelines for analyzing bias and propaganda in English texts. There are totally 8 bias labels and 4 propaganda labels.
- Task 2: Bias and Propaganda annotation in Tamil - Three language policy Dataset.
  The goal of task 2 is to provide annotating content related to the Three Language Policy/India's National Education Policy related issues. The task is to annotate based on the bias and propaganda guidelines in Tamil texts that discuss this policy. There are totally 7 bias labels and 4 propaganda labels.

Annotation Guidelines

- **Unbiased:** Neutral / Without favoritism.
  Example: "The US Gender Policy aims to address the rights of transgender individuals in military service, but the policy has been met with mixed reactions from different communities."
- **Biased Against US Gender Policy:** Criticizes negatively.

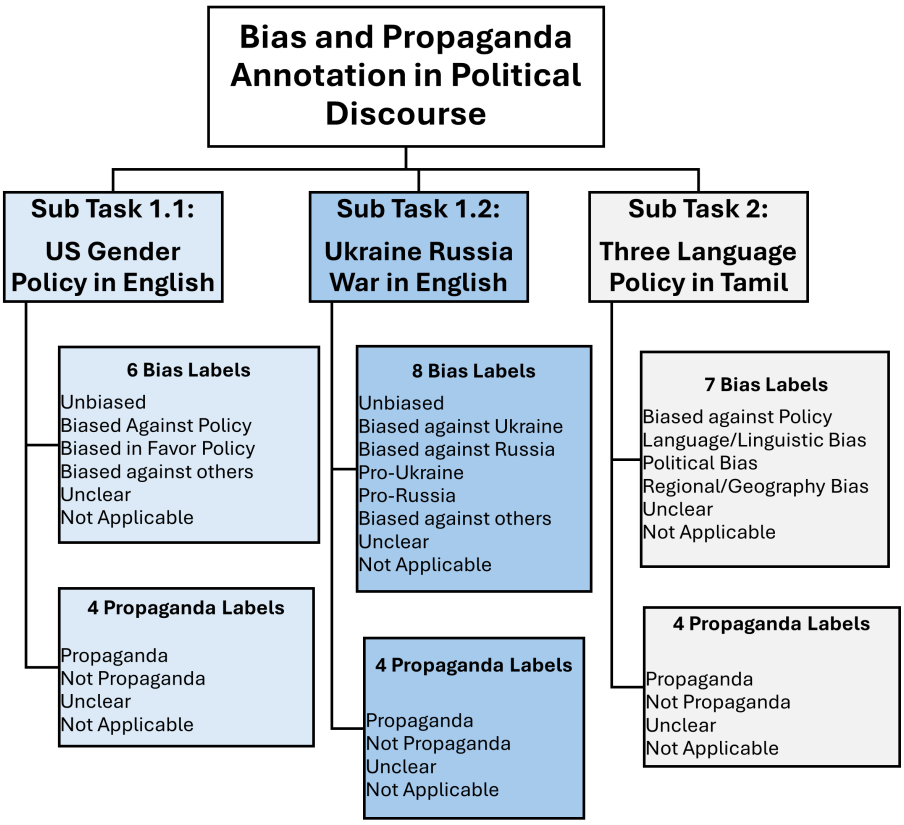

Figure 1: Task Overview and Annotation Labels

Example: "The US Gender Policy discriminates against transgender individuals by denying them the right to serve in the military, an unjust decision that harms the LGBTQ+ community."

- **Biased in Favor of US Gender Policy:** Strongly supports.
  Example: "Trump's US Gender Policy is a necessary measure to protect national security and uphold traditional values, and it's a step in the right direction for the country."
- **Biased Against Others:** Criticizes others negatively.
  Example: "Those who oppose Trump's US Gender Policy are out of touch with reality and are prioritizing political correctness over national security."
- **Unclear:** Text is incomplete.
  Example: "The debate over the US Gender Policy continues, but many people are still unsure about its long-term impact."
- **Not Applicable:** Irrelevant to the topic.
  Example: "The latest economic report shows growth in GDP this quarter."

4 Dataset Description

The dataset has been carefully curated from the YouTube platform by collecting comments from videos discussing the political concern. To identify relevant videos, we utilized a combination of hashtags such as *'National Education Policy'*, *'Three Language Policy'*, *'US gender policy'*, *'Ukraine Russia War'* alongside manual keyword searches including terms like *'Zelensky'*, *'Putin'* and *'Trump US Gender Policy'*.

After gathering the comments, we applied preprocessing steps to remove unrelated or noisy content. This included removing usernames, URLs, and comments containing fewer than three words to ensure data quality and relevance. The dataset statistics is shown in table 1

5 Participant Methodology

A total of 20 teams registered to participate in this shared task. However, only 2 teams submitted their

Task	No. of Samples	Vocab Size	Avg Length (in tokens)
Sub Task 1.1	7,911	9,475	21.54
Sub Task 1.2	5,099	12,619	37.95
Task 2	5,880	33,076	28.97

Table 1: Dataset statistics across different subtasks

results. Following are the detailed methodology of the participating teams.

- Scalar: Team Scalar contributed to Subtask 1.1 by performing manual annotations on the provided textual data, focusing on identifying bias and propaganda techniques present in discourse related to US gender policy. The annotation was carried out by two undergraduate students, both aged between 20–23. The manual annotation effort by Team Scalar is critical in generating high-quality, labeled datasets for training robust NLP models.

  Team Scalar also developed a transformer-based NLP model to detect both propaganda techniques and bias using the annotated data. They trained the model using transfer learning on a specially annotated dataset label encoding bias categories and tagging six propaganda techniques while adding extra non-propaganda examples to reduce class imbalance. The model was optimized with Adam and trained for four epochs (batch size 32) using sparse categorical crossentropy, achieving roughly 47.9 % accuracy.

- Mithun: This team present a context-aware neural model for detecting bias and propaganda in multilingual political discourse. Their approach stands out for its comprehensive annotation methodology, leveraging advanced metrics such as Bias Score, Cosine Similarity, Fairness Difference, and Weighted F1-score to evaluate both the fairness and accuracy of language models across diverse demographic and linguistic groups. By applying these metrics to English and Tamil datasets on sensitive topics like US gender policy, Ukraine/Russia discourse, and Three Language Policy the participant demonstrate significant disparities in model performance and fairness, highlighting the persistent challenges of bias in multilingual NLP.

6 Results and Discussion

Team Name (Sub Task 1.1)	Cohen's Kappa	Rank
Scalar	0.71	1
Mithun	0.39	2

Sub Task 1.2	Rank
Mithun (0.42)	1

Task 2: Tamil	Rank
Mithun (0.48)	1

Table 2: Bias and Propaganda Annotation Task results across English and Tamil subtasks.

Evaluation Metric: Cohen's Kappa

Cohen's Kappa (κ) is a statistical measure used to assess inter-annotator agreement for categorical classification tasks while correcting for chance agreement. It is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Where:

- P_o is the observed agreement between annotators,
- P_e is the expected agreement by random chance.

A κ value of:

- 1 indicates perfect agreement,
- 0 indicates chance-level agreement,
- Negative values indicate systematic disagreement.

Results and Quantitative Analysis

Table 2 presents the performance outcomes of participating systems across multiple subtasks:

- **Sub Task 1.1 (English)**: Team *Scalar* achieved the highest Kappa score of 0.71, indicating substantial agreement and strong classification capability. Team *Mithun* followed with 0.39, reflecting moderate agreement and highlighting difficulties in capturing nuanced propaganda techniques in English.
- **Sub Task 1.2 (English)**: Despite a moderate Kappa score of 0.42, *Mithun* ranked first, implying effective relative performance on this subtask.

- **Task 2 (Tamil)**: *Mithun* attained a Kappa of 0.48, leading the task. This score reflects moderate agreement in a low-resource language scenario, where annotation and detection challenges are more pronounced.

7 Conclusion

This shared task aims to enhance the annotation process of bias and propaganda in multilingual political discourse, focusing on English and Tamil texts. The shared task highlighted the role of clear guidelines, examples, and collaboration in advancing NLP research on complex, sensitive, and opinion analysis tasks. The resulting dataset and insights contribute valuable resources and direction for future work in this important area. Despite limited submissions, the task underscored the challenges in multilingual annotation and the importance of culturally-informed guidelines. Future efforts will focus on expanding the dataset, refining the annotation schema, and encouraging broader participation to build more generalizable models.

Acknowledgments

This work was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight).

References

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. 2021. Fine-grained classification of political bias in German news: A data set and initial experiments. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131, Online. Association for Computational Linguistics.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. 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)*, pages 4982–4991, Online. Association for Computational Linguistics.

Adam J Berinsky. 2017. Rumors and health care reform: Experiments in political misinformation. *British journal of political science*, 47(2):241–262.

Marco Casavantes, Manuel Montes-y Gómez, Luis Carlos González, and Alberto Barróñ-Cedeno. 2024. Propitter: A twitter corpus for computational propaganda detection. In *Advances in Soft Computing*, pages 16–27, Cham. Springer Nature Switzerland.

Tanmay Chavan and Aditya Manish Kane. 2022. ChavanKane at WANLP 2022 shared task: Large language models for multi-label propaganda detection. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 515–519, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Pratim Datta, Mark Whitmore, and Joseph K. Nwankpa. 2021. A perfect storm: Social media news, psychological biases, and ai. *Digital Threats*, 2(2).

Andrew M. Guess and Benjamin A. Lyons. 2020. *Misinformation, Disinformation, and Online Propaganda*, page 10–33. SSRC Anxieties of Democracy. Cambridge University Press.

Freddy Heppell, Kalina Bontcheva, and Carolina Scarton. 2023. Analysing state-backed propaganda websites: a new dataset and linguistic study. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5729–5741, Singapore. Association for Computational Linguistics.

Sora Lim, Adam Jatowt, Michael Färber, and Masatoshi Yoshikawa. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1478–1484, Marseille, France. European Language Resources Association.

Ankita Maity, Anubhav Sharma, Rudra Dhar, Tushar Abhishek, Manish Gupta, and Vasudeva Varma. 2024. Multilingual bias detection and mitigation for Indian languages. In *Proceedings of the 7th Workshop on Indian Language Data: Resources and Evaluation*, pages 24–29, Torino, Italia. ELRA and ICCL.

Preslav Nakov and Giovanni Da San Martino. 2020. Fact-checking, fake news, propaganda, and media bias: Truth seeking in the post-truth era. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 7–19, Online. Association for Computational Linguistics.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017*

Conference on Empirical Methods in Natural Language Processing, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Bonnie Steinbock. 1978. Speciesism and the idea of equality. *Philosophy*, 53(204):247–256.

Matthew Welsh and S. Begg. 2016. What have we learnt? insights from a decade of bias research. *Australian Petroleum Production and Exploration Association Journal*, 56.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa El-Beltagy, and Muhammed AbuOdeh. 2024. The FIGNEWS shared task on news media narratives. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 530–547, Bangkok, Thailand. Association for Computational Linguistics.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11(3):e0150989.