Bias Bluff Busters at FIGNEWS 2024 Shared Task: Enhancing Bias Detection through Annotator Awareness and AI Integration

Silvia Pareti[1]* **Jasmin Heierli**[2]* **Serena Pareti**[3] **Tatiana Lando**[1]
[1]Independent Researcher [2]Zurich University of Applied Sciences
[3]Università Cattolica del Sacro Cuore
si.pareti@gmail.com heej@zhaw.ch
serena.pareti01@icatt.it tatiana.lando@gmail.com

Abstract

This paper details our participation in the FIGNEWS-2024 shared task on bias and propaganda annotation in Gaza conflict news. Our objectives were to develop robust guidelines and annotate a substantial dataset to enhance bias detection. We focused on creating guidelines that prompt annotators to be aware of their biases and emphasise the need for a diverse annotator pool, to ensure that high inter-annotator agreement (IAA) reflects genuine consensus rather than homogeneity. We iteratively refined our guidelines and included detailed examples for clarity. We also explored the integration of ChatGPT as an annotator to support consistency. Our findings include the challenges in eliciting annotation consistency without embedding bias in the guidelines themselves and the importance of fostering annotator awareness of their own biases. This work provides insights into the complexities of subjective bias annotation and offers well-crafted guidelines for the field.

1 Introduction

Media bias detection is a rapidly evolving field dedicated to developing tools that help news producers, analysts, and readers identify and mitigate biased content. Such biases can profoundly shape public perception and understanding of events. By improving transparency and promoting objectivity, these tools aim to foster a more informed and balanced media landscape, ensuring that diverse perspectives are fairly represented. However, the field faces significant challenges. Data scarcity, lack of standard definitions for bias and annotation (Spinde et al., 2024), and generally low annotator agreement hinder the development of effective bias detection models. Existing large-scale datasets typically contain one to two thousand sentences annotated by experts or micro-jobbers (Färber et al., 2020; Spinde, 2021; Kiesel et al., 2019).

The FIGNEWS-2024 shared task (Zaghouani et al., 2024) is a competition aimed at annotating bias and propaganda in news and developing guidelines to support such annotation. Related shared tasks or datathons, like SEMEVAL-2019 (Kiesel et al., 2019) or "Hack the News Datathon"[1] provided news dataset already annotated with bias or propaganda, and participants competed on their automatic detection. FIGNEWS-2024 task focuses instead on the annotation itself of bias and propaganda in news related to the conflict in Gaza, considering as primary targets the entities of Palestine and Israel. The goal of the task is to develop guidelines and annotate each text with labels indicating the presence or absence of negative bias or propaganda towards a target entity (see Table 2 in A).

The task texts vary in type and length, from 5 to over 1400 words, including news articles and social media texts, which may contain titles, article bodies, links, and hashtags. Only negative bias is annotated, with Israel and Palestine as the primary targets; any other entity falls under 'bias against others'. The dataset was split into two sections: 'Main', which required 1800 texts (Batch1 and 2) to be at least single-annotated, and 'IAA', which required 200 texts from the same batches to be independently annotated by all participants.

Our submission aligns with the shared task requirements and goals. Our contributions include the creation of detailed guidelines for text-level annotation of bias and propaganda and the annotation of 2,000 texts, achieving Krippendorff's alpha agreements of 0.43 (Kappa 43.3) and 0.3 (Kappa 31.5) respectively.

*Equal contribution.

[1]https://www.datasciencesociety.net/hack-news-datathon/

Proceedings of The Second Arabic Natural Language Processing Conference, pages 580–589

2 Annotation Methodology and Examples

2.1 Development of Annotation Guidelines

We developed the guidelines using a lightweight process suited to the approximately one-month time frame and participants' availability. The development was structured in the following stages:

1. **DRAFT:** Conducted literature research and trial annotation on 20 texts to familiarize with the data and discuss alignment. Drafted initial guidelines (v.0.1).

2. **DEVELOPMENT:** Annotated 1800 texts from the Main dataset, collected examples, and discussed edge cases. Updated guidelines with examples, decisions, and edge cases.

3. **COMPLETION:** Finalized guidelines to version 1.0 before annotating the IAA datasets.

During annotation development, we defined in-scope bias types and analyzed news language peculiarities, such as quotations and clickbait. We clarified the relationship between bias and propaganda, identifying linguistic clues and indicators from Recasens et al. (2013) for bias and Da San Martino et al. (2019) and Piskorski et al. (2023) for propaganda, adding more as encountered in the data. We also established clear definitions for the referents of Palestine and Israel and developed a glossary of common terms used to refer to entities and events, aligning on them being expressions of positive, neutral, or negative bias.

2.2 Data Annotation Process

Before starting the annotations, annotators participated in an exercise to discuss their assumptions and reflect on their own biases. They discussed these biases as a group to raise awareness and mitigate their impact on the annotation process.

Besides an initial alignment set of 20 examples, all annotation was performed independently by the annotators. For the text in the Main section of the dataset, annotators could consult one another and collect edge cases for group discussion. Annotators were asked to read the texts and formulate a justification for their chosen label before applying it. This process helped reduce the effect of their unconscious bias by ensuring they could ground their decisions in the guidelines. Annotators could consult the guidelines to check the definition for each label, the list of possible cues for bias, the suggested connotation for prominent words, and the conventions for annotating quotations, clickbait, and other special cases peculiar to news. We also formulated a set of tests to support the decisions:

- *Could you rewrite the text to make it more neutral?*
- *Reading this text, could you tell who the reporter supports among the parties involved?*
- *Would you still consider this biased/unbiased if the original text was in Arabic or Hebrew?*
- *Is the text de-/humanizing one side?*

Other resources annotators could use included conducting searches to make more informed decisions about the factuality of events or any intentional information omissions and consulting ChatGPT's annotations of bias and propaganda and connotated words (see Sec.2.2.1).

2.2.1 ChatGPT as an Annotator

For the shared task, we experimented with integrating ChatGPT[2] as one of the annotators for the Main dataset section. Each annotation was performed using a separate prompt to ensure accuracy and context-specific analysis.

We provided the LLM with the labels definition for the task and instructed it to add a label as well as highlight any connotated words in the text. Our goals were two-fold: 1) Support the annotators' work by helping them identify potential bias-carrying words. 2) Improve annotation quality by creating a two-way annotation system, allowing us to review and arbitrate cases where the human annotator and ChatGPT disagreed.

2.3 Inter-Annotator Agreement (IAA) Analysis

Inter-annotator agreement (IAA) on bias annotation tasks is typically low, due to the subjective nature of the task. For instance, Spinde (2021) reports a Fleiss' Kappa of 0.21 for sentence-level bias annotation, and Hube and Fetahu (2019) notes an α agreement of 0.124. Although annotators in these studies were not experts, their task consisted only in determining the presence or not of bias. Our annotation was carried on by experts, but it required the application of multiple labels (see Table 2 in A). We achieved α=0.43 (Kappa 43.3) for the bias task and α=0.30 (Kappa 31.5) for the propaganda task, indicating fair to moderate agreement. Full

[2] https://github.com/zhaw-iwi/FIGNEWS-2024

results and ranking we achieved for each subtask are summarized in Appendix D[3].

These results demonstrate that even with a team of experts and detailed guidelines, bias annotation remains highly subjective. Increased alignment and training could help annotators achieve better convergence, but the submission deadline was a limiting factor.

2.3.1 Agreement with ChatGPT

Additionally, we analyzed the inter-annotator agreement between our human annotators and ChatGPT for bias and propaganda annotation over texts in the Main section of the dataset. Initially, the Krippendorff's α for bias annotation was 0.143. However, after cleaning the data to account for excusable errors, the score improved significantly. The cleaning process involved removing:

- 64 instances labeled as "unbiased against someone," an invented label.
- One response, "please provide me with the text to analyse," which indicated a misunderstanding by ChatGPT.
- 788 instances labeled as "biased against Palestine," due to misinterpretation of Hamas as synonymous with Palestine, which we clarify in our guidelines, but was not specified in the prompt.

Post-cleaning, the Krippendorff's α for bias annotation improved to 0.315, which is comparable to the agreement levels achieved between the annotation team members. For propaganda annotation, the Krippendorff's α between ChatGPT and human annotators was 0.346. These results demonstrate that ChatGPT can achieve a level of agreement comparable to human annotators in both bias and propaganda tasks, reinforcing the potential of using LLMs to support human annotation efforts with proper prompting.

3 Team Composition and Training

The team (Table. 1) comprises four female members residing in three different countries in Western Europe. Two annotators are native Italian speakers, while the others are native German and Russian speakers, respectively. All members are fluent in English and hold MA, BA, or PhD degrees in Linguistics and Computational Linguistics. Three annotators have no stake in the conflict or direct links to the involved parties or religions, whereas one annotator is originally from Eastern Europe and of Jewish ethnicity.

The team met weekly online to plan the work, discuss annotation cases, make guideline decisions, and coordinate. To familiarize themselves with the task, the annotators independently tagged 20 texts, followed by thorough discussions of the findings and uncertainties, which informed the initial drafting of the guidelines. During the annotation of the Main section of the dataset, annotators reported ambiguous or edge case texts in a discussion document to ensure continuous alignment. Communication was maintained through a group chat platform, enabling prompt feedback exchanges and fostering collaboration among team members throughout the project duration.

4 Task Participation and Results

Our team focused on developing robust guidelines, prioritizing the discussion of annotation challenges, documenting decisions on edge cases, and deriving general principles to help annotators recognize their own biases and make more grounded decisions. Due to time constraints, we acknowledged the limitation in providing extensive training and alignment opportunities for better annotator convergence.

One key takeaway from participating in the task was the realization that bias is inherent in human nature and in human-created products, such as ChatGPT. Consequently, some level of bias in the annotators is unavoidable and will inevitably affect annotation decisions. A good practice, therefore, would be to have a large and diverse group of annotators to represent multiple perspectives. Nonetheless, we found that explicitly identifying bias and propaganda indicators fostered a deeper sense of awareness and helped annotators make more informed decisions.

The team annotated one-way the 1800 text from the Main section of the dataset and 4-way 200 texts from the IAA section. Looking at the Main section, the labels assigned by the annotators show a skewed distribution. Fig. 1 and 2 in C show that roughly one third of texts were considered 'Unbiased' and another third as expressing 'Bias against others', largely referring to Hamas. Texts biased against Israel and Palestine represented around 10% and 2.5% respectively. We also identified over 21% 'Unclear' cases, partly covering texts expressing

[3]The full results across all teams can be found in Zaghouani et al. (2024).

	Native	Gender	Age	Region	Education	Expertise
Ann1	Italian	F	34-44	Western Europe	PhD	Linguistics, NLP
Ann2	German	F	34-44	Western Europe	MA	NLP
Ann3	Russian	F	34-44	Western Europe	MA	NLP
Ann4	Italian	F	18-24	Western Europe	BA	CL

Table 1: Team demographics.

positive bias. Table 5 shows the labels distribution across bias and propaganda, showing clear correlations.

5 Discussion

5.1 Contributions

Our team's participation in FIGNEWS-2024 has led to the following contributions to the field of bias and propaganda annotation:

- **Detailed guidelines** for annotating news and quoted text with examples and edge cases.
- **Conscious bias methodology** to increase annotators' own bias awareness.
- **Entity definition** of what constitutes Palestine and Israel (see Table 3), and guidelines for making context-aware decisions for entities that could refer to these.
- **Connotation glossary** of common terms used to refer to events and entities in the conflict (see Table 4).
- **Test questions** to help annotators detect bias.
- Identified a **new category of persuasive strategies for propaganda**, termed 'Humanization,' which involves adding unnecessary details, such as personal stories of victims, to establish a connection between the reader and the depicted party.
- Clarified the **relationship between bias and propaganda**, noting that propaganda is itself an indicator of intentional bias.

5.2 Recommendations

Based on our findings, we recommend the following for a bias annotation task:

- **Introduce Labels for Positive Bias:** Including labels for positive bias will provide a more comprehensive analysis of the text, capturing both favorable and unfavorable biases.
- **Split the Labeling Process:** Divide the labeling process into several steps to enhance accuracy and clarity:
  1. Bias detection.
  2. Polarity tagging (positive/ negative).
  3. Target entity identification.
  4. Optionally, tag the intensity of the bias.
- **Allow for Multi-layered Labeling:** Implement multi-layered labeling to account for different voices in the text, such as the reporter and quoted sources. This approach will help in distinguishing the biases present in various narrative elements.
- **N-way Annotation:** Implement N-way annotation for all examples. Although time-consuming, this approach enhances accuracy for complex tasks and helps identify tags that frequently cause confusion.

5.3 Other observations

During the task, we encountered several issues with the data:

- **Lack of Context:** The absence of full article context might affect judgment, as it is unclear how the text is framed within the entire article.
- **Issues with Translated Texts:** The intermediate translation step can change the connotation and perception of the texts. Moreover, readers are unlikely to naturally consume automatically translated texts.
- **Disclosure of Original Language:** Disclosing the original language of the text can bias annotators. We still decided to disclose it because it proved critical for understanding the potential relationship between the reporter and the entities, leading to more accurate interpretations.

6 Conclusion

In this paper, we have detailed our approach and findings from participating in the FIGNEWS-2024 shared task, focusing on the annotation of bias and propaganda in news texts. Our contributions include the development of comprehensive guidelines that emphasize the need for annotator awareness of their own biases and the integration of linguistic clues to detect bias and propaganda. Despite

achieving moderate inter-annotator agreement, our work highlights the inherent subjectivity in bias annotation and the necessity for extensive training and alignment among annotators. Our findings suggest that bias annotation guidelines[4] should be tested with a diverse pool of annotator to ensure high IAA is the product of quality guidelines rather then of annotators' homogeneity.

Future work should consider introducing labels for positive bias, refining annotation processes to account for different voices within texts, and addressing the challenges posed by translated texts and lack of context. This study underscores the complexity of media bias detection and the ongoing need for refined methodologies and diverse perspectives in the annotation process.

References

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. 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.

Michael Färber, Victoria Burkard, Adam Jatowt, and Sora Lim. 2020. A multidimensional dataset based on crowdsourcing for analyzing and detecting news bias. In *Proceedings of CIKM 2020*, CIKM '20, page 3007–3014, New York, NY, USA. Association for Computing Machinery.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, page 195–203, New York, NY, USA. Association for Computing Machinery.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, et al. 2023. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, Technical report, European Commission Joint Research Centre, Ispra (Italy).

[4] https://github.com/zhaw-iwi/FIGNEWS-2024

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.

Timo Spinde. 2021. An interdisciplinary approach for the automated detection and visualization of media bias in news articles. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 1096–1103.

Timo Spinde, Smi Hinterreiter, Fabian Haak, Terry Ruas, Helge Giese, Norman Meuschke, and Bela Gipp. 2024. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *Preprint*, arXiv:2312.16148.

Wajdi Zaghouani, Mustafa Jarrar, Nizar Habash, Houda Bouamor, Imed Zitouni, Mona Diab, Samhaa R. El-Beltagy, and Muhammed AbuOdeh, editors. 2024. *The FIGNEWS Shared Task on News Media Narratives*. Association for Computational Linguistics, Bangkok, Thailand.

A Set of Labels

	Bias Annotation	Propaganda Annotation
Labels	Unbiased, Biased against Palestine, Biased against Israel, Biased against both Palestine and Israel, Biased against others, Unclear, Not Applicable	Propaganda, Not Propaganda, Unclear, Not Applicable

Table 2: The full set of labels given for each of the tasks.

B Treatment of Specific Terminologies and References

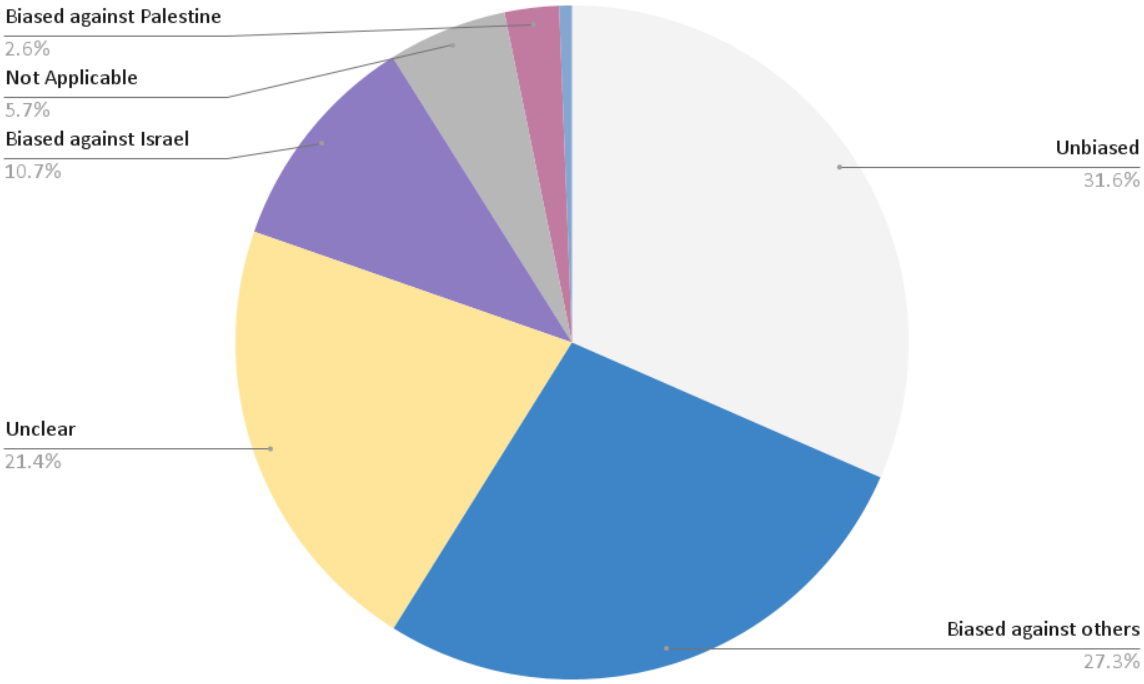
	Refers to Palestine?	Refers to Israel?
Yes	• Muslims/ Arabs (Palestinian majority also Muslim) in the context of the current war • Gaza/ West Bank (although we recognize that they have important differences)	• Jews (Israel majority also Jews) • Israel's army/troops/IDF
No	• Hamas: not equated with Palestine. Mentions of Hamas are related to Palestine only if explicitly stated • ISIS: not equated with Hamas or Palestine. Mentions of ISIS are related to Palestine only if explicitly stated • Palestinian Islamic Jihad: not Palestine nor Hamas • Palestinian authority, politicians, representatives (Mohammad Mustafa), unless used to refer to Palestine as a whole	• Zionists: not equated with Jews or Israel. Mentions of Zionists are related to Israel only if explicitly stated • Specific parties or prominents figures within Israel politicians or army, including the president Netanyahu, settlers/settlements, political parties and representatives.

Table 3: Definition of what might constitute a reference to Palestine or Israel. Note: this was provided to the annotators for consistency. It is the product of the authors' own considerations and might reflect their own bias.

Entities/ events	Expresses bias against entity	Expresses bias for entity	Not bias
Palestine	Hamas, Isis	Palestine country	Gaza, Palestine, Palestine territories
Israel	Occupation army, occupation government, Zionist entity, Settlers, Israeli army	-	Israel, Israeli government, Israeli forces
October 7th (Hamas)	-	Response/ reaction, military operation	Attack, offensive, armed incursion, invasion, operation
War (Israel)	genocide, Israel's revenge, aggression, massacre, crimes against humanity, war crimes	self-defense	Israel-Hamas (unless used to suggest it is only Israel's war)
Hamas	Terrorists, monsters, murderers	Resistance, liberation organization, movement	Militants, fighters, organization
Casualties	Assassination, murder, massacre	-	Death, killing
Hamas hostages	-	-	Hostages, kidnapped, captives, Detainees, prisoners
Israeli prisoners	Hostages, kidnapped	-	Prisoners, detainees

Table 4: Definition of common references encountered for relevant entities and events in the task with the authors' decision about these carrying a positive or negative connotation. Note: this was provided to the annotators for consistency. It is the product of the authors' own considerations and might reflect their own bias.

C Label Distribution Across the Main Dataset



Bias	Propaganda NA	Not Propaganda	Propaganda	Unclear	Grand Total
Biased against both		5	6		11
Biased against Israel		54	125	14	193
Biased against others		123	349	20	492
Biased against Palestine		3	42	2	47
Not Applicable	93	3	5	3	104
Unbiased		559	2	6	567
Unclear		117	117	152	386
Grand Total	93	864	646	197	1800

Table 5: Labels distributions on the Main section of the dataset.

Figure 1: Labels distribution for bias over the 1800 texts annotated in the Main section of the dataset.

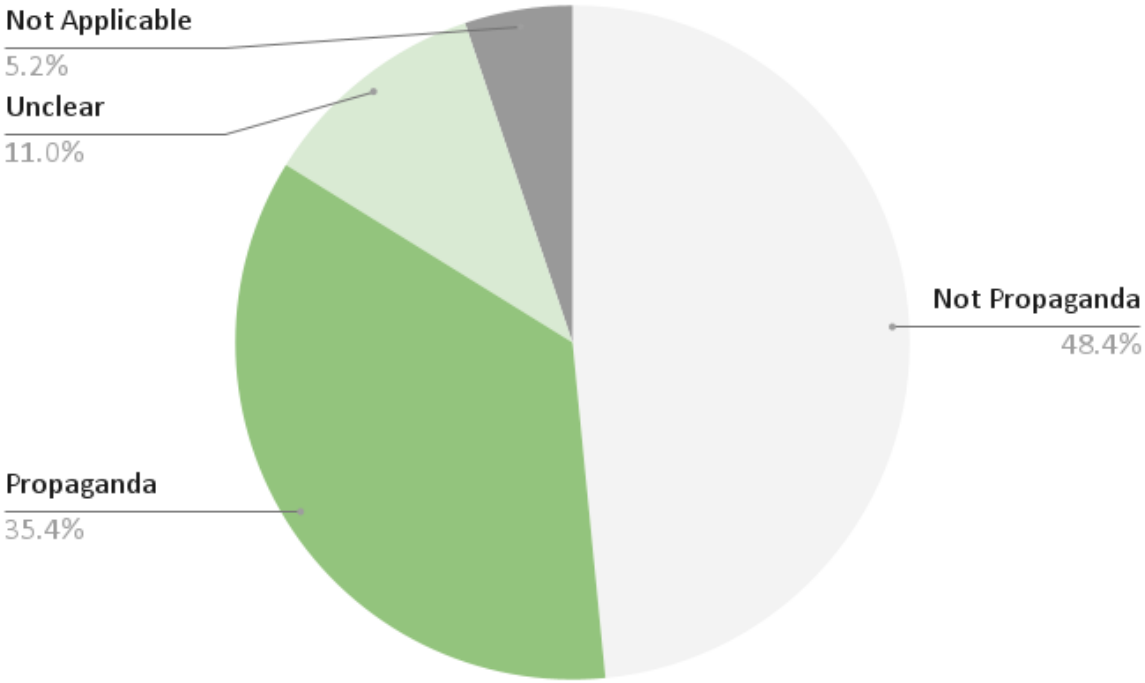


Figure 2: Labels distribution for propaganda over the 1800 texts annotated in the Main section of the dataset.

D Full Team Annotation Stats

Bias Metrics	Quantity (Data Points)	Quality (Kappa)	Centrality (Macro F1 Avg)
Results	2,600	43.3	21.0
Avg all teams	5,531	44.5	24.7
Rank	9	9	13

Table 6: Team results for the Bias subtask. 16 teams participated in this subtask. Centrality refers to the Highest Cross-Team Macro F1 Average of Bias Labels on both B01 and B02 batches

Propaganda Metrics	Quantity (Data Points)	Quality (Kappa)	Centrality (Macro F1 Avg)
Results	2,600	31.5	**37.6**
Avg all teams	6,883	33.5	37.4
Rank	4	4	**2**

Table 7: Team results for the Propaganda subtask. 6 teams participated in this subtask. Centrality refers to the Highest Cross-Team Macro F1 Average of Bias Labels on both B01 and B02 batches

E All teams vs BBB Team Labels Distribution

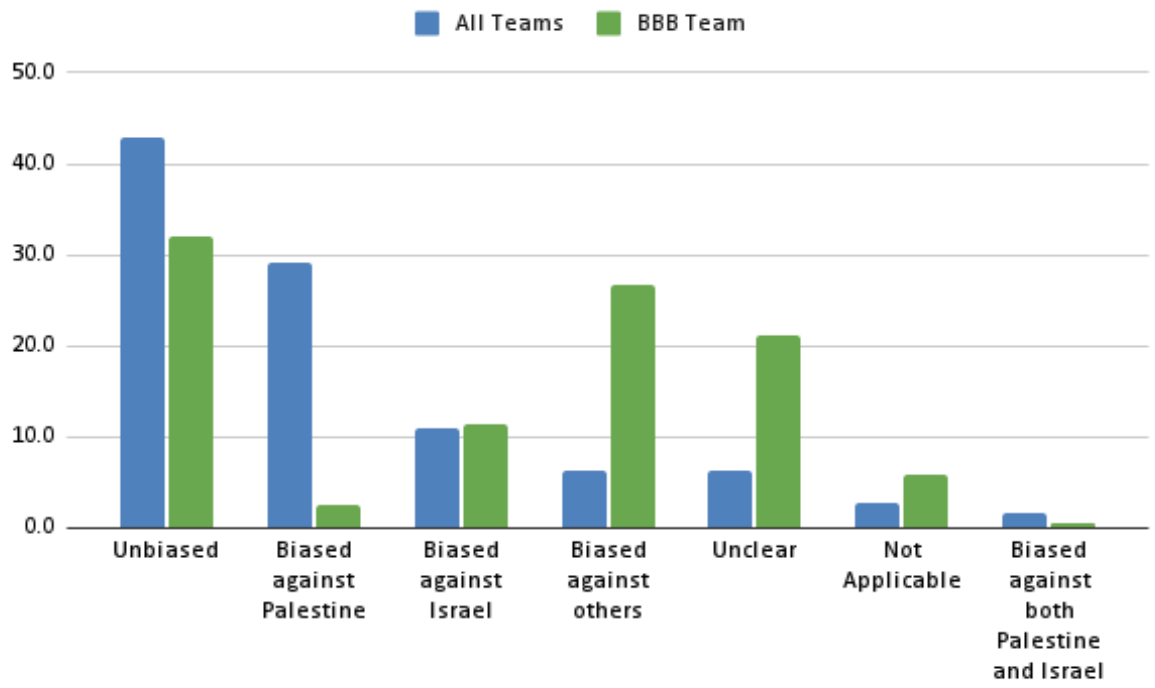


Figure 3: Labels distribution across all teams vs our team: Bias Bluff Busters (BBB). This counts all labels given by the annotators across Main and IAA sections of the dataset. The discrepancy in distribution can be explained by our team decision to treat Hamas not as equivalent to Palestine and annotate it as 'Bias against others' and by treating positive bias as expressing bias with 'Unclear' target.

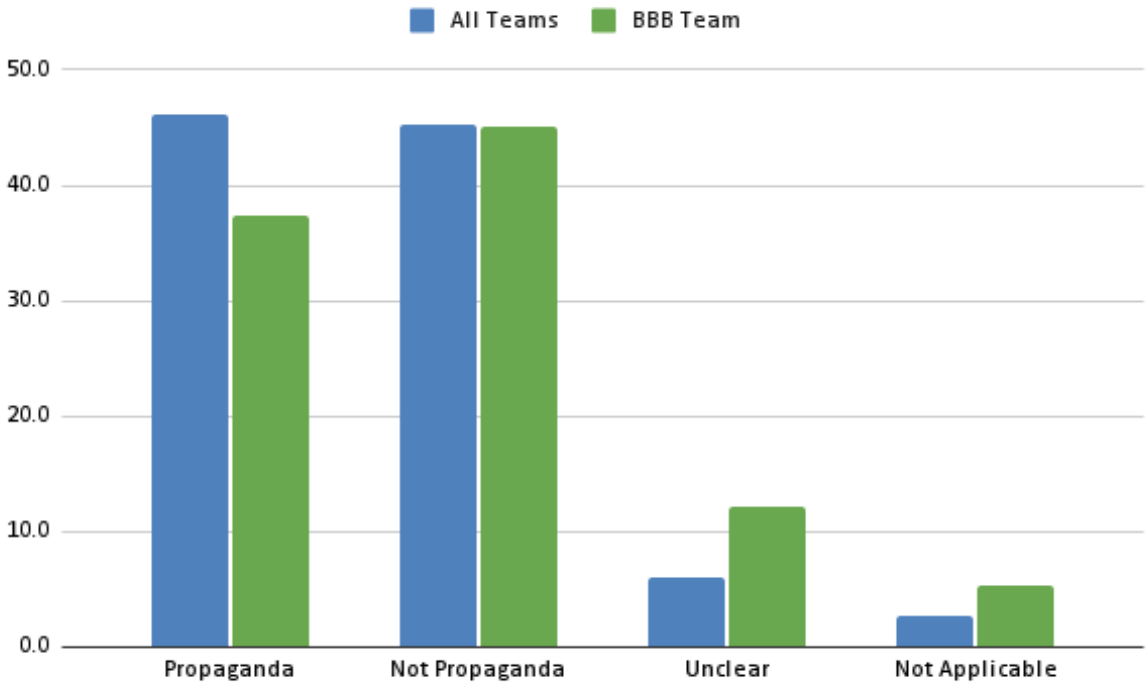


Figure 4: Labels distribution across all teams vs our team: Bias Bluff Busters (BBB). This counts all labels given by the annotators across Main and IAA sections of the dataset.