BSC-LANGTECH at FIGNEWS 2024 Shared Task: Exploring Semi-Automatic Bias Annotation using Frame Analysis

Valle Ruiz-Fernández*, José Javier Saiz*, Aitor Gonzalez-Agirre
Barcelona Supercomputing Center
{valle.ruizfernandez,jose.saiz,aitor.gonzalez}@bsc.es

Abstract

This paper introduces the methodology of BSC-LANGTECH team for the FIGNEWS 2024 Shared Task on News Media Narratives. Following the bias annotation subtask, we apply the theory and methods of framing analysis to develop guidelines to annotate bias in the corpus provided by the task organizators. The manual annotation of a subset, with which a moderate IAA agreement has been achieved, is further used in Deep Learning techniques to explore automatic annotation and test the reliability of our framework.

1 Introduction

Being the first and primary source of information, online news articles play a key role in how people shape their opinions and engage with topics of societal relevance (Hamborg et al., 2019). However, media coverage is often far from being unbiased. Considering the impact of media bias on public opinion, Zaghouani et al. (2024) propose Framing the Israel War on Gaza: A Shared Task on News Media Narratives (FIGNEWS 2024). This shared task focuses on exploring the media narratives and digging deep into the biases and propaganda in news articles about the Gaza-Israel 2023-2024 war from multiple perspectives, cultures and languages (Arabic, Hebrew, Hindi, French and English). The main objective is to develop guidelines to annotate a corpus of news posts. Such corpus was compiled from international news article headlines and advertising posts from Facebook in the languages above mentioned. The posts selected date from October 1, 2023 to January 31, 2024, and they all include the word query "Gaza" in any of the different languages. In addition, the corpus also contains machine translations into English and Arabic for all posts. With this aim, this shared task proposed two possible subtasks to focus: annotating either bias or propaganda. As the title suggests, our work uses practical conceptualisations from framing theory to define each label and constrain corner cases within the bias annotation task; and leverages encoder-decoder classification models to aid the annotation process, while also being useful to evaluate the robustness of the proposed annotation framework within the automatic modelling of media framing.

Efforts to annotate and detect bias have been already made (Recasens et al., 2013; Cremisini et al., 2019; Baly et al., 2020), some of them focusing on the Palestinian-Israeli conflict (Lin et al., 2006; Greene and Resnik, 2009; Al-Sarraj and Lubbad, 2018). In addition, some shared tasks have also been devoted to delve into bias in news articles, like the CLEF-2023 CheckThat! Shared Task (Da San Martino et al., 2023). However, most of these works do not include detailed information about the annotation, and the guidelines used are not publicly available neither.

In this work, we apply the theory and methods of framing analysis to develop guidelines for bias annotation in a specific topic. Specifically, we combine linguistic frame indicators from Recasens et al. (2013) and narrative framing from Frermann et al. (2023) to solve the bias detection task as a single-label classification problem.

Computational approaches in framing device annotation are numerous, but neural network models have been the most popular approach for the most recent years, including BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) classification models (Ali and Hassan, 2022). For this reason, we decide to explore automatic multi-label prediction of our frames using the base version of RoBERTa model (Liu et al., 2019), which allows us to test the reliability of our framework.

The paper follows with Section 2, which explains the guideline creation and data annotation process, and discusses the application of the proposed framework to the dataset. Section 3 introduces the team of annotators and the quality check procedures used to ensure consistency and expectations between annotators. Section 4 discusses the Inter-Annotator Agreement (IAA) results and the experiments with the supervised prediction of our labels.

*Equal contribution

Proceedings of The Second Arabic Natural Language Processing Conference, pages 620–629
August 16, 2024 ©2024 Association for Computational Linguistics

2 Annotation Methodology and Examples

2.1 Development of Annotation Guidelines

Zaghouani et al. (2024) proposed the following seven classes to annotate bias in the FIGNEWS corpus: *Unbiased*, *Biased against Palestine*, *Biased against Israel*, *Biased against both Palestine and Israel*, *Biased against others*, *Unclear* and *Not Applicable*. Given these labels, the annotation guidelines were designed based on the three main steps into which we divided the annotation process: (1) Determining the applicability of the post, (2) determining the existence of bias, and (3) deciding the bias direction. Appendix A contains the annotation guidelines, and Figure 1 in the same appendix shows the decision tree diagram including these three main steps.

Determining the applicability of the post. All posts that do not mention and do not contain information about Israel conflict in Gaza are considered as *Not Applicable* and, thus, not taken into account in the remaining annotation process.

Determining the existence of bias. Given that journalistic texts are not supposed to be opinionated, we follow the linguistic indicators proposed by Recasens et al. (2013) to consider whether a piece is biased or not, which involve identifying implicit sentiments or perspectives based on lexical choices. Furthermore, while media frames can be analysed at the linguistic level, as suggested by Recasens et al. (2013), they can also be analysed at the narrative level, which implies multiple levels of framing, including the co-occurrence of frames and narrative roles. This is in contrast to focusing only on localised signals, as in existing NLP approaches. Following Frermann et al. (2023), we use high-level framing devices that are established in the communication theory (Semetko and Valkenburg, 2000). We adopt the narrative frames applicable in our topic as indicators of bias, and use the questions from Frermann et al. (2023) to define each frame.

Deciding the bias direction. If a post is considered biased, in order to develop our guidelines on how to determine the direction of bias, we no longer focus on lexical indicators, but on syntax (Greene and Resnik, 2009) to identify the key actors and, in particular, following the work of Frermann et al. (2023). Apart from the narrative frames that were discussed before, we apply three narrative roles to analyze the implication of each actor in the story: "hero", the actor who contributes to/is responsible for solving the problem; "villain", the actor who contributes to/is responsible for causing the problem, and "victim", the actor who suffers the consequences of the problem. The bias direction is determined taking into account the respective narrative roles of the actors in each post.

While our annotation guidelines aimed to be clear, unambiguous and comprehensive from the very beginning, we organized multiple calibration sessions to ensure consistency, as specified in Section 3. In our guidelines, we also provide clarification for the ambiguous cases that have emerged along the three calibration sessions organised.

2.2 Data Annotation Process

The corpus provided for the shared task annotation is organized in 15 batches of 1,000 post/batch. For each batch, a 10% (100 post/batch) is dedicated to the inter-annotator agreement (IAA) analysis. For this work, we have only considered the posts in the first two batches (B01, B02). The data annotation process involved three steps: (1) manual annotation of the posts later used to calculate the IAA, (2) fine-tuning two models to automatically annotate the remaining posts, and (3) manual revision of the automatic annotation.

Manual annotation. The guidelines designed were used to annotate a subset of batches B01 and B02 later used to calculate the IAA. Results are provided in Section 4.

Model fine-tuning[1]. To test the scalability and application of our annotation framework, we have explored automatic annotation using a transformer-based classification model, and use it to annotate the remaining posts. After manually labelling the posts in the IAA section, the English machine translation version of these posts and their bias labels were used as training data for the fine-tuning. Note that posts with lower annotator agreement were removed, i.e. only posts where both annotators agreed on the same label were kept.

[1] Fine-tuning a pre-trained large language model involves optimising its parameters to improve performance on a particular task. This process involves augmenting the model's training data with task-specific examples. In our case, we focus on text classification, aiming to refine the model's ability to categorise posts based on their exhibited biases.

The data was divided into three subsets: training (70%), evaluation (15%) and test (15%), with an equal distribution of labels. We performed two fine-tunings of the base version of RoBERTa (Liu et al., 2019), a large pre-trained language model already trained on a large number of corpora. Our goal was to obtain two models that could later be used to predict the bias label of the sentences: `model_1` predicts whether the post is biased or not. If the resulting label is *Biased*, then `model_2` is used to predict the direction of bias and thus the final label.

Both models were fine-tuned for 10 epochs with a maximum sequence length of 128, a batch size of 32 and a learning rate set at 5e-5. All model implementations, along with the code for fine-tuning and evaluation, are sourced from the Hugging Face's Transformers library[2] (Wolf et al., 2020). We performed 4 runs with different random seeds for both fine-tunings. As shown in Table 1, with the best seed, `model_1` achieved a F1 score of 0.59 on the test set, and `model_2`, 0.82.

Manual revision of automatic annotation. Both resulting models have been used to annotate the remaining posts of batches B01 and B02. However, note that the predictions and the probability score of the label in each case have been only used as a guidance. This automatic annotation has been further revised by the same humans in charge of the annotation task, taking into account its complexity. Once again, this final annotation have been carried out considering only the English machine translation of the posts. The results of comparing the automatic annotation with the final label after the human revision are displayed in Table 1.

2.3 Inter-Annotator Agreement (IAA) Analysis

The IAA was calculated with 200 posts, corresponding to a 10% each batch B01 and B02, labeled individually by each annotator. The results show a Kappa coefficient of 0.51. The overall accuracy achieved was 65.5%, with a macro F1 average of 39.8%, reflecting a significant difference in performance between the classes. The F1 Bias[3] was 81.5%, showing that the guidelines proved effective in detecting whether a post was biased or not.

[2]https://github.com/huggingface/transformers

[3]F1 Bias is the F1 of the group label Bias, which equates all forms of bias labels as a single class.

3 Team Composition and Training

Annotator demographics. Both annotators in charge of the whole annotation task, a man and a woman, are also the two main authors of this paper. We fall within the age range of 18 to 24 years old, and we share a cultural, linguistic and educational background. Coming from Spain, our native language is Spanish, and we both have a Translation and Linguistics background, now working in the area of NLP.

Team coordination. Both annotators individually annotated the 200 items reserved for the IAA calculation, and the revision of the final labels of the remaining instances was divided equally: one batch per annotator. During the annotation process, we organised three calibration sessions between the two annotators to discuss ambiguous cases. Although these sessions were also used to refine the annotation guidelines, the aim was to exchange feedback and provide general guidance to ensure consistency between the annotators.

Training. As mentioned, the annotators in charge of the whole annotation task are also the main authors of this paper and the guidelines proposed. Thus, annotators were fully familiar with the annotation task and also with the context of the Israel conflict in Gaza. Given this background knowledge, no specific training was deemed necessary prior to starting the annotation process.

However, to ensure consistency and reliability in the annotations, before the individual annotation of the posts set aside for the IAA, both annotators carried out individual annotations of 75 instances taken from the remaining 90% of the posts in batch B01 as a test. Annotations were then shared and discussed. This preliminary phase served to:

1. Check the applicability of the guidelines, identify any ambiguities or challenges in the instructions and address them collaboratively.
2. Discuss and solve disagreements to align interpretations and approaches, ensuring a more unified understanding and application of the guidelines.
3. Build confidence in the task by familiarizing annotators with the nuances of the annotation process.

Despite the annotators' familiarity with the task and context, identifying and interpreting a frame is highly dependent on the contextual information available to the annotator (Baden, 2018; D'Angelo, 2018). Therefore, in scenarios where annotators might lack familiarity with the task or topic, additional training and support should be considered. Such training could include detailed briefings on the subject matter, guided annotation exercises with feedback, and iterative discussions. This would help ensure that all annotators have a robust and consistent understanding, thereby enhancing the reliability and validity of the annotations.

4 Task Participation and Results

Our team processed a total of 2,200 instances across 2 main batches, where the IAA subset of each batch, corresponding to 200 out of 2000 instances, was annotated individually by each annotator.

Centrality measures show agreement across different teams in both the main and IAA annotations, where our Kappa ranked 5th place with a 26.5%. Our accuracy ranked 1st place with a 46.6%, and a Macro F1 average of 29.2% in 5th place, and an F1 bias of 60.2% in the 9th place. These centrality measures highlight a high overall accuracy, while also shows the challenges in creating annotations that are consistent across other guidelines.

Regarding the automatic annotation, Table 1 shows the performance of the models used to annotate 90% of the posts in B01 and B02 with respect to the human revision. Regarding *Biased*, *Unbiased* and *Biased against Palestine* classes, the human agreed with the label predicted by the models in almost 70% of the cases during the review process, and thus no further changes were done. Note that, in the case of the *Biased against Israel* class, the percentage of agreement increases by about 15%.

However, the overall agreement between the automatic and human annotations is lower (39.27% for `model_1` and 62.83% for `model_2`) due to the complete lack of agreement in the remaining classes, showing that the models were not able to learn them. This could be due to the small amount of data on which the model was trained and its unbalanced label distribution, where for `model_1` the label *Not Applicable* was significantly less present than *Biased* and *Unbiased*. Similarly, in the training data of `model_2`, *Biased against both*, *Biased against others* and *Unclear* are less prominent than the others.

	`model_1`	`model_2`
F1	**0.59**	**0.82**
% agreement w/ human ann.:		
Not Applicable	0.00%	-
Unbiased	68.56%	-
Biased	68.34%	-
Biased against Palestine	-	68.60%
Biased against Israel	-	83.00%
Biased against both	-	0.00%
Biased against others	-	0.00%
Unclear	-	0.00%
total	**39.27%**	**62.83%**

Table 1: F1 score on the test set for the best seed of `model_1` and `model_2`, and percentage of agreement with the final label after human revision of the automatic annotation.

5 Discussion

Our goal was to define annotation guidelines for identifying media frames that can be used to detect bias as a single-label classification problem. To this end, we adapt a set of linguistic and narrative frame indicators from the discussed literature into a deductive framework. We then apply the framework to annotate the FIGNEWS 2024 corpus, which contains news articles about the Gaza-Israel war of 2023-2024. Our proposed framework was used to annotate and explore supervised Deep Learning techniques using RoBERTa models, which showed a higher correlation with human annotation in the most prominent labels of the training dataset.

However, our contribution faces some limitations: the annotation task was fully carried out by the same people developing the guidelines. Introducing some external annotators not familiar in the task would be helpful to get a deeper insight of the explainability and consistency of the guidelines.

Regarding the IAA results, our annotations achieved only a moderate agreement in the IAA analysis. A high level of agreement can be challenging in such complex annotation tasks, but our result suggests a meaningful alignment in the interpretations of the annotation guidelines. Moving forward, we can further refine our guidelines and potentially enhance agreement in future annotations.

In addition, among the 15 batches available, only the first 2 of them have been annotated and used in this work. Having limited data has also constrained the results of our fine-tunings. Having a larger amount of training data would improve the performance of the models. In relation with the fine-tuning experiments carried out, we could explore further the impact of training with a different set of hyperparameters, models or even training data scenarios.

6 Conclusion

Our work contributes to the creation of detailed, consistent guidelines considering both theory and methods of framing analysis for a complex annotation task such as the presented in the FIGNEWS 2024 shared task on News Media Narratives. More specifically, we have developed guidelines to annotate bias on news posts about the Gaza-Israel 2023-2024 war. In addition, we have explored state-of-the-art Deep Learning techniques to explore the automatization of the annotation task to easen it.

Acknowledgements

This work has been promoted and financed by the Government of Catalonia through the Aina project. This work is funded by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335, 2022/TL22/00215334. We thank the collaboration of Lidia Neves, from the Federal University of Espírito Santo.

References

Wael F Al-Sarraj and Heba M Lubbad. 2018. Bias detection of palestinian/israeli conflict in western media: A sentiment analysis experimental study. In *2018 International Conference on Promising Electronic Technologies (ICPET)*, pages 98–103. IEEE.

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christian Baden. 2018. *Reconstructing Frames from Intertextual News Discourse*, page 3–26. Routledge.

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.

Andres Cremisini, Daniela Aguilar, and Mark A Finlayson. 2019. A challenging dataset for bias detection: the case of the crisis in the ukraine. In *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12*, pages 173–183. Springer.

Giovanni Da San Martino, Firoj Alam, Maram Hasanain, Rabindra Nath Nandi, Dilshod Azizov, and Preslav Nakov. 2023. Overview of the clef-2023 checkthat! lab task 3 on political bias of news articles and news media. *Working Notes of CLEF*.

Paul D'Angelo, editor. 2018. *Doing news framing analysis II*. Routledge, London, England.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander 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, New York City. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Holli A. Semetko and Patti M. Valkenburg Valkenburg. 2000. Framing european politics: A content analysis of press and television news. *Journal of Communication*, 50(2):93–109.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

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 Annotation Guidelines

This appendix presents the bias annotation guidelines in the dataset proposed in the FIGNEWS 2024 Shared Task on News Media Narratives (Zaghouani et al., 2024). These annotation guidelines propose a method to identify which side of the war the bias of each post is against and formulates the problem as a text classification task. Section A.1 describes the annotation classes used and the steps and guidelines followed to label the posts, section A.2 is devoted to solve the ambiguities annotators may encounter, and, finally, Section A.3 summarizes some ethical considerations.

A.1 Annotation Process

Posts are labelled following the 7 classes proposed by Zaghouani et al. (2024):

- *Unbiased*
- *Biased against Palestine*
- *Biased against Israel*
- *Biased against both Palestine and Israel*
- *Biased against others*
- *Unclear*
- *Not Applicable*

The annotation process defined here involves three steps: (1) Determining the applicability of the post (Section A.1.1), (2) determining the existence of bias (Section A.1.2), and (3) deciding the bias direction (Section A.1.3). Figure 1 shows the decision tree diagram including these three steps. Note that all posts are annotated taking into account the English machine translation.

A.1.1 Determining the applicability of the post

Before determining whether a post is biased against a certain side or not, all posts that do not mention and do not contain information about Israel conflict in Gaza are annotated with the label *Not Applicable*. For example:

> When freedom of speech becomes freedom of hate it becomes a whole different thing.

For all the remaining posts, which do contain information about Gaza war, the label is decided depending on whether they show bias or not, as described in Section A.1.2.

A.1.2 Determining the existence of bias

Given a post not annotated as *Not Applicable*, it is considered to be biased if it contains, at least, one of the following features:

- **Factive verbs**, which pressupose the truth of the complement they introduce (Recasens et al., 2013).

  > Hero: The medic who entered Lutofat in Kibbutz Kfar Gaza in an unprotected ambulance recalls - "I **realized** that people will continue to die".

  Note, however, that the same verb can be used in a non-factive way (for example, if the verb presents someone's opinion or an experimental result) (Recasens et al., 2013).

  > Intelligence documents seized by the IDF during the fighting in Gaza **reveal** that terrorist organizations use mosques in the Strip for terrorist purposes.

  Also, even if the verb is used as a factive, it is not an indicator of bias itself if the agent of such factive verb is a neutral entity (e.g. NGOs, international political entities, financial entities).

  > "Gentlemen, it is also important to **realize** that the "Hamas" attacks did not occur in a vacuum" Statements of the Secretary-General of the United Nations, António Guterres. [...]

- **Entailments**, i.e. words or phrases whose truth implies the truth of another one. In the following example, "murder" imply "kill" in an unlawful, cruel way (Recasens et al., 2013).

  > In 26 days, Israel **massacred** more than 9000 civilians in Gaza. More than 3000 innocent children were **brutally murdered**. This is an act of terror, ethnic cleansing, genocide. Israel is terrorist.

  The use of "murder" in the post above contrasts with the use of "kill" in the following one, in which there is no entailment:

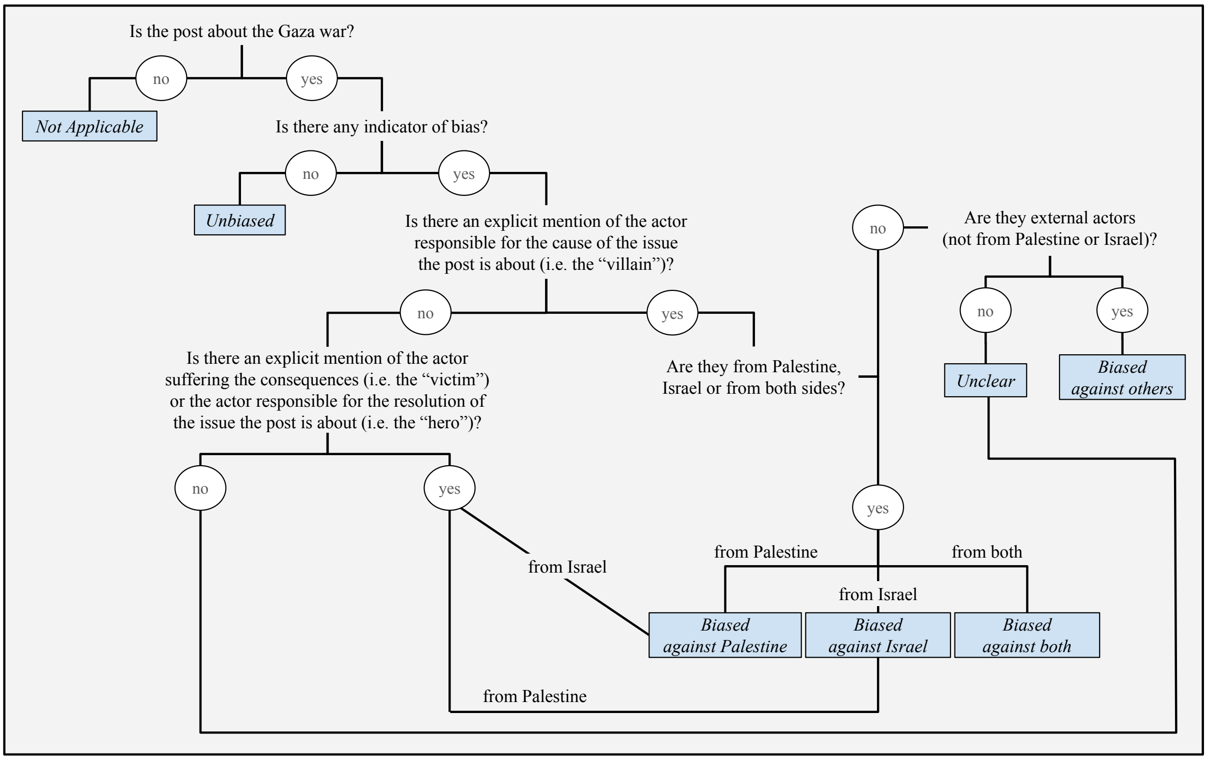

Figure 1: Decision tree diagram followed to annotate bias in the FIGNEWS corpus.

> Hamas says a top official was among four people **killed** in an explosion in Beirut.

- **Assertive verbs**, which cast doubts on the proposition they introduce. In contrast to factive verbs, assertive verbs do not directly presupose the truth of the proposition; however, they imply a certain level of certainty (varying depending on the verb) (Recasens et al., 2013).

> People who **claim** that Hamas is a resistance movement are either hypocritical, naive, anti-Semitic, or all three.

- **Subjective terms or intensifiers**, i.e. adjectives or adverbs that add a subjective emphasis to the sentence (Recasens et al., 2013).

> British journalist Yotam Confino watched video clips collected from ISIS Hamas cameras on October 7 and recounts the **horrific atrocities** he witnessed. [...]

- **One-side terms** reflecting only one of the sides of the war (Recasens et al., 2013).

> International forces are welcome if they want to **liberate** #Palestine. [...]

- **Metaphors and comparisons** used to equate entities or realities with a negative connotation, such as hell or the evil, with one of the sides of the war.

> This morning we saw the **face of evil**. Hamas launched a criminal attack, without distinguishing between women, children and the elderly. [...]

- **Human Interest binary questions** that indicate whether or not the post puts a human face on the conflict, employing personal testimonies or other linguistic resources that may generate strong feelings on the reader. Adapted from Semetko and Valkenburg (2000).

> Hapoel "Shlomo" Tel Aviv participates **with great sorrow and deep pain in the mourning of the Liebstein family** for the death of the team’s fan, Nitzan, who was murdered in his kibbutz, Kfar Gaza. [...]

- **Moral binary questions** that indicate whether or not the post contains a moral message or makes reference to morality or religious tenets. Adapted from Semetko and Valkenburg (2000).

> [...] The world must understand what happened here, that the people of Hamas murdered In cold blood 600 merciless Israelis, **may their name and memory be blessed**, the world must denounce them and immediately!

If a post does not contain any of the items listed, it is considered that it only contains factual information, even if it is negative and/or it harms the image of the agent or side. Thus, it will be classified as *Unbiased*. Otherwise, the post will be considered biased and further labeled specifying the side it is against following the criteria in Section A.1.3.

A.1.3 Deciding the bias direction

If a post is considered biased according to section A.1.2, it is then analyzed how framing turns bias against a particular side. To do this, the first step is to identify the key actors in the narrative. Note that an actor in this context can be a public figure (e.g. a politician), but also organisations, groups or the civil population. The posts can show one of the following syntactical forms used to identify the key actors of the narrative:

- **3-way transitive relations**, where both the subject agent and the direct object are present in the sentence (Greene and Resnik, 2009).

  > On October 7, the members of the Nazi Hamas slaughtered a woman who had helped the residents of Gaza all her life, and they reached her with accurate information and brutally murdered her, until it took 40 days to identify the body. [. . .]

- **Nominalized forms** as subjects omitting the agent (Greene and Resnik, 2009).

  > Urgent | The assassination of the Deputy Head of the Political Bureau of Hamas, Saleh Al-Arouri, [. . .]

- **Passive forms** omitting the agent (Greene and Resnik, 2009).

  > [. . .] Ofir Liebstein, head of the Shaar Negev Regional Council, who was murdered today while heroically defending his residents.

Once the key actors in the post have been identified, we assign their corresponding narrative roles between those that are responsible for the issue ("villains"), those who are affected ("victims") and those who can resolve the issue ("hero"), as adopted in Frermann et al. (2023).

If the subject agent of the post who is responsible for an issue, i.e. the "villain" of the narrative, is from Israel or Palestine, the post is labelled against one of them. Otherwise, if the actor is not directly from one of these territories, it is considered an external actor. Depending on which side the villain belongs to, posts will be labelled according to the criteria below:

- *Biased against both*: Both Israel and Palestine are framed as "villain".

  > Since we witness the eradication of a people under the pretext of obeying the urgency of saving another, is it even possible to hope for a part of humanism, of respect for life? , in the face of this savage persecution which falls every day, incessantly, without restraint, without conscience on Gaza. [...]

- *Biased against Israel*: Israel is framed as "villain".

  > Violent Israeli bombardment on the Gaza Strip [...].

  Note that, when both an Israel-based actor and an external actor is framed as "villain" in the same post, the overall frame is considered to be against Israel. For example:

  > Faced with the ongoing humanitarian catastrophe in Gaza, you should be ashamed, Madam President of the European Commission. Shame on turning a blind eye to the deliberate massacre of civilians. Ashamed of supporting Netanyahu when already 4,000 children have died under his bombs. Shame on refusing the ceasefire!

- *Biased against Palestine*: Palestine is framed as "villain".

  > On 7.10.23 - They came to brutally attack the towns of southern Israel in an attempt to extinguish the light of a people who wrote heroic epics throughout history. 7.12.23 - We brought them with our lights to the depths of their existence to rid the world of their darkness and the blackness of their hearts between the two dates - Message: The victory of light over darkness represented by Hamas terrorists ISIS

  Once again, when both a Palestine-based actor and an external actor is framed as "villain" in the same post, the overall frame is considered to be against Palestine. For example:

  > US President Joe Biden says both Hamas and Russia's Vladimir Putin "want to completely annihilate a neighboring democracy."

- *Biased against others*: Only one or multiple external actors are framed as "villains".

> Yemen's Houthis have waded into the Israel-Hamas war raging more than 1,000 miles from their seat of power in Sanaa, declaring they fired drones and missiles at Israel in attacks that highlight the regional risks of the conflict.

If the post explicitly mentions only one actor as either "hero" or "victim", this means that it is presented in a positive frame. In this case, we turn to the deductive process, where the other side of the war is implicitly portrayed in a negative frame. For example, if Israel attacks are being portrayed as outstanding achievements in the war, the post is implicitly framing Palestine as the "villain", as in the following example. Therefore, the post will be biased against the opposite side of the war, which is not mentioned explicitly in the post.

> A sad Hanukkah holiday. Hanukkah, the holiday of heroism. Two more heroic reservists of the people of Israel fell today in the Gaza Strip: Major Gal Meir Eisenkot, 25 years old from Herzliya, son of the former Chief of Staff and member of the cabinet Gadi Eisenkot, and Sergeant Major Yonatan David Dietsch, 34 years old from Mahrish. Their deaths will not be in vain. May their memory be blessed!

Finally, if a post is biased but the "villain" actor is not specified or ambiguous, it is considered *Unclear*, as in the following example:

> And I crumpled to the ground and I said, 'They are in the midst of killing our son.

A.2 Handling Ambiguities and Consistency

Ambiguity can arise from multiple interpretations of language and different perspectives, especially in complex opinion annotation tasks such as bias detection. While our annotation guidelines aim to be clear and comprehensive to minimise this problem, we have also organised review sessions to ensure consistency among annotators and gather feedback to refine the guidelines. More specifically, these review sessions are designed to flag ambiguous cases, discuss them within the proposed annotation framework, and update the annotation guidelines if a case cannot be resolved. We provide clarification for the ambiguous cases that have emerged consistently along the 3 review sessions we have organised:

- **Civilians express an opinion against the combatants/politicians of their same territory**, in which case the negative frame is against the corresponding side of the war. For example:

  > Gaza: "The people want to overthrow Hamas" Gazans chant this Saturday in the corridor which connects Khan Younès to the protected humanitarian zone.

- **Calls for action** are moral messages that create social expectations on how to behave, and are considered to be biased. For example:

  > [...] The campaign is still long, and we are expected to have a continuous and stubborn fight against a barbaric and bloodthirsty enemy who seeks to destroy us. We must show patience, maintain national resilience, remain united, and hold our heads high as much as possible. We will forever remember our heroic soldiers who did not hesitate to enter the cursed land of terror to destroy evil and thus protect us all. May their memory be blessed and enshrined in the heart of the nation forever.

- **Emojis and hashtags** should be also considered in the annotation, as they provide information of the author's opinion on the conflict, even when the main body of the post itself is not clear on the topic or the author's position. For example:

  > A Controversial Video Goes Viral on Social Media: Women Laugh and Take Selfies in the Background of a Kidnapping, Displaying Disrespectful Gestures. #Israel #HamasWar **#IsraelUnderAttack** #Gaza #Palestinians #IsraelPalestineWar #Gaza **#IsraelFightsTerror #IndiaStandWithIsrael** #telaviv #Hezbollah #FPJ

- **The Israel War on Gaza is mentioned but is not the main topic of the post**, and there is not enough information to identify bias frames directly related to the war. In this case, the post will be labeled as *Unbiased*. The label *Not applicable* should be only used to label only those posts that do not mention the Israel War on Gaza. For example:

  > Date 17 October, day Tuesday, this BBC Hindi podcast is full news all day long. Today in the program, we talked about the Supreme Court's decision on gay marriage and the **ongoing conflict between Israel and Hamas in Gaza** [...].

- **The Israel War on Gaza is critiziced but no side is mentioned.** Therefore, the responsibility of the conflict is assumed to be placed on both sides of the conflict, and thus the post is labeled as *Biased against both*. For example:

  > "Hell on earth" for the UN official on site, "carnage" for UNICEF, "intolerable human suffering" for the Red Cross. It's been 7 days since the truce ended. [...]

A.3 Ethical Considerations

When annotating bias in news posts, it is crucial to apply these guidelines, but also to take into account various ethical considerations to ensure fairness, accuracy, and transparency. On the one hand, annotators should strive to remain as objective as possible, setting aside personal beliefs and opinions. Annotators should prevent their own biases from influencing the task and consider the different viewpoints a news post may present. On the other hand, we are aware that we are working with sensitive data that may be considered offensive. This data must be used only for the specific purpose of annotating bias and exploring media narratives.