FIGNEWS: Bias Annotation Guidelines

Valle Ruiz-Fernández* and **José Javier Saiz***
{valle.ruizfernandez, jose.saiz}@bsc.es
Barcelona Supercomputing Center

1 Introduction

This document presents the **bias annotation guidelines** in the dataset proposed in the FIGNEWS Shared Task on News Media Narratives (Zaghouani et al., 2024). The objective is to detect and classify bias in a corpus of news posts about the Palestine-Israel War on Gaza. More specifically, the corpus was compiled from international news article headlines and advertising posts from Facebook in English, Arabic, Hebrew, French, and Hindi. The posts selected date from October 1, 2023 to January 31, 2024, and they all include the word query "Gaza" in one of the languages mentioned. In addition, the corpus also contains machine translations into English and Arabic for all posts.

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 2 describes the annotation classes used and the steps and guidelines followed to label the posts. Section 3 is devoted to solve the ambiguities annotators may encounter. Section 4 presents the tools used to carry out the annotation, and, finally, Section 5 summarizes some ethical considerations.

2 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 2.1), (2) determining the existence of bias (Section 2.2), and (3) deciding the bias direction (Section 2.3). Figure 1 shows the decision tree diagram including these three steps. Note that all posts have been annotated taking into account the English machine translation.

2.1 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-Palestine conflict 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 2.2.

2.2 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.

*Equal contribution

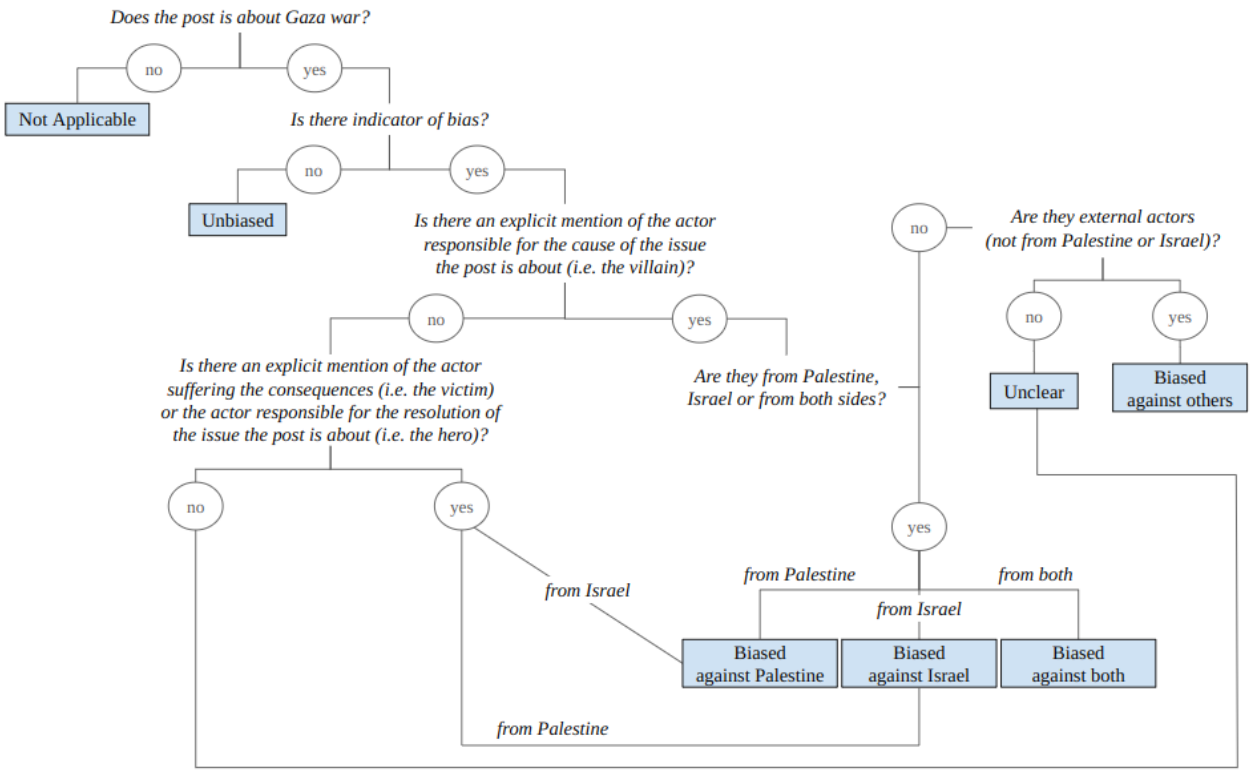

Figure 1: Decision tree diagram followed to annotate bias in the FIGNEWS corpus.

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:

> 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 connota-

tion, 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 2.3.

2.3 Deciding the bias direction

If a post is considered biased according to section 2.2, 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.

3 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. [...]

4 Annotation Tools

Posts are annotated directly in the Google sheet provided by the organizators of the FIGNEWS Shared Task (Zaghouani et al., 2024). However, we have also made use of Deep Learning techniques to easen the annotation task. More specifically, we have performed two fine-tunings of roberta-base (Liu et al., 2019), a large pre-trained language model already trained on gigantic corpora. Fine-tuning this model involves further training it using a domain-specific dataset to solve a specific task —in our case, text classification—. Our aim was to obtain two models that could be later used to predict the bias label of the posts in the MAIN section of the Google sheet provided: the first model predicts if the post is biased or unbiased. If the resulting label is *Biased*, the second model is then used to predict the bias direction and, thus, the final label.

All model implementations, along with the code for fine-tuning and evaluation, are sourced from the Hugging Face's Transformers library[1] (Wolf et al., 2020). In our case, these fine-tunings of roberta-base were performed after labeling manually the posts in the IAA section of the Google sheet, following the guidelines designed. Once annotated, the English machine-translation version of such posts and their bias label were used as the data for the fine-tunings. This data was split into 3 subsets: train (70%), evaluation (30%) and test (30%), with equal distribution. We performed 4 restarts with different random seeds for both fine-tunings. With the best seed, the first model achieved a f1 score of 0.59 on the test set, and the second one, 0.82.

As mentioned, both models have been used to annotate the posts in the MAIN section. 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 needs to be always further revised by the humans in charge of the annotation task, taking into account its complexity.

5 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 opin-

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

ions. 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 only used for the specific purpose of annotating bias and exploring media narratives.

References

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.

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.

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 Raed AbuOdeh, editors. 2024. *The FIGNEWS Shared Task on News Media Narratives*. Association for Computational Linguistics, Bangkok, Thailand.