The CyberEquity Lab at FIGNEWS 2024 Shared Task: Annotating a Corpus of Facebook Posts to Label Bias and Propaganda in Gaza-Israel War Coverage in Five Languages

Mohammed H.S. Helal, Radi Jarrar,
Mohammad Alkhanafseh, Abdallah Karakra, Ruba Awadallah

Department of Computer Science, Birzeit University, Palestine

{mhelal,rjarrar,malkhanafseh,akarakra,rawadallah}@birzeit.edu

Abstract

This paper presents The_CyberEquity_Lab team's participation in the FIGNEWS 2024 Shared Task (Zaghouani, et al., 2024). The task is to annotate a corpus of Facebook posts into bias and propaganda in covering the Gaza-Israel war. The posts represent news articles written in five different languages. The paper presents the guidelines of annotation that the team has adhered in identifying both bias and propaganda in coverage of this continuous conflict.

1 Introduction

The ongoing conflict between Palestinians and Israelis has been a subject to traditional media bias and propaganda for years, layers over layers of misinformation accumulated over decades did not help bridge the gap between Palestinian and Israeli public points of views. Then comes the event of October 7th, coinciding with the era of Social Media, where people can see and hear the suffering of innocent war victims through numerous social platforms, alternative to traditional media outlets. One would think that in the age of information, controlling a narrative would be out of the reach, however since the events of 7th of October and followed by Israel's military response, social media has been plagued with bias and propaganda, and the same narratives and talking points from media, globally, are echoing all over social networks.

Detecting bias and propaganda from news or social media posts has been deeply explored in the literature and a variety of methodologies and techniques have been presented. For instance, detecting bias through sentiment analysis on news reports using Natural Language Processing (NLP) and Machine Learning (ML) techniques (Park, 2009). Hutto *et. al*. presented sentiment analysis, subjectivity analysis, modality and other analysis to detect bias in short sentences (Hutto, 2015). Shahi *et. al.* has presented a semi-automatic annotation framework to detect misinformation on COVID-19 pandemic in social media posts (Shahi, 2022). Abuaiadah *et. al.* performed Clustering on Arabic Tweets in order to detect sentiment (D. Abuaiadah, 2017).

Generating ML models, through NLP techniques, rely on large corpora for training to achieve satisfying level of accuracy. ML models learn from such corpora and use its knowledge to perform numerous operations on new data. In order to facilitate the use of such models to detect bias and propaganda in the Israel-Gaza social media coverage, trained annotation teams are required to label a large portion of the corpora while following well-defined labeling guidelines. This research work has been done as participation in FIGNEWS 2024 shared task at the ArabicNLP 2024 conference. The task aims at constructing a corpus of Facebook posts and crowdsourcing an annotation outline to label bias and propaganda in these posts (Zaghouani, et al., 2024). The task is split into two subtasks one for labeling bias and the other is for labeling propaganda. The corpus consists of 15000 posts collected from October 1st, 2023 till January 31st, 2024 and written in five different languages: Arabic, Hebrew, French, Hindi, and English. The posts are translated into Arabic and English using translation tools. The corpus was collected by searching for the keyword "Gaza" in the five aforementioned languages.

In this paper, The CyberEquity Lab proposes two sets of guidelines to define and identify both bias and propaganda in the coverage of Israel-Gaza war. The rest of the paper is organized as follows: section 2 will discuss annotation process and methodology, section 3 will present team

Proceedings of The Second Arabic Natural Language Processing Conference, pages 614–619
August 16, 2024 ©2024 Association for Computational Linguistics

composition and training. Team participation and results will be discussed in section 4. Finally, discussion and conclusion will be presented in sections 5 and 6.

2 Annotation Methodology

Labeling bias and propaganda in full news articles is has been always a challenge. However, what is more challenging is when labeling in short social media posts, some might be as short as a single sentence. In order to develop a comprehensive set of guidelines for such challenging annotation task, our team had to go through several revisions and endless debates, scratching our head: how to define bias and propaganda, what is the reference, what is the yardstick. Williams (Williams, 1975) defined bias as willful, influential, and threatening to widely held conventions. To apply Williams's definition on our task, one must ask an important question: which conventions are we talking about, Palestinian or Israeli? The next section will discuss the development process that led our team to the concluded annotation guidelines.

2.1 Development of Annotation Guidelines

FAIR (FAIR.org) presents a set of guidelines on how to generally identify bias, the proposed guidelines are adopted by many institutions including Lehman Collage and it is summarized by answering these questions:

1. Who is reporting?
2. Who is funding?
3. What are the unchallenged assumptions and stereotypes?
4. Is there loaded language?
5. Does the story present false balance between the two sides?

The list provided by FAIR is not applicable on short Facebook texts. However, a general understanding of the definition of bias began to crystalize through multiple discussions between our team members. For practical reasons (such as time-efficiency and error reduction), the guidelines need to be short and precise. Here is the proposed guidelines for both bias and propaganda, each in its own sub-section.

2.1.1 Guideline for Classifying Bias

1. **Dehumanization of a group**. When a group of people are being attacked for the sole sake of belonging to a particular group identity, or when a group is demonized to justify violations of their basic human rights. For example, the post "*There are voices in Israel calling for the destruction of all buildings in the Gaza Strip, not as revenge but as a tactical solution*…" refers to the destruction of Gaza strip, in which millions of Palestinians live in.
2. **Hate Speech**. Oxford Language Dictionary's definition is "abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds." An example from our corpus: "*Palestine..Pakistan..'Hamas' brothers of both sold the material to shed blood to Hamas..this is how Pakistan is.*" The post indicates hate speech and bias against Palestinians and Pakistanis although nothing seems to be in common between the two nations except for their religion.
3. **Double Standards**. Holding two different groups onto different sets of standards, which implies favoritism and bias towards one side over the other. For example, the following post was posted while thousands of Palestinian children are being subjected to violence and starvation: "*World's saddest birthday: Kfir Bibas marks first birthday in Hamas captivity*".
4. **Misinformation.** Targeting a group of people with smearing via misinformation.
5. **Labeling and Name Calling.** Giving offensive labels for a particular group with the intention to alienate them and therefor dehumanize them. For example, the following post referring to bombing of The Baptist Hospital in Gaza: "*Outrage and violence erupted in response to media reports that Israel hit a Gaza hospital in an airstrike, but Israel has presented evidence that the explosion was actually caused by a terrorist rocket that misfired.*" Israel is accused of targeting that hospital and media outlets adopt Israel narrative without providing any evidence. Later, the narrative was debunked from media. The Baptist hospital is one of 30

other hospital Israel has bombed so far. Calling Hamas terrorists and accepting Israel's narrative is bias.

2.1.2 Guideline for Classifying Propaganda

The guideline for annotating propaganda is adopted mainly from (Da San Martino, *et al*., 2019) (Dimitrov, *et al.*, 2021) (Alam *et al.*, 2022). We selected 12 categories from the list of 20 categories. These categories belong to four main classes: Appeal to Commonality (such as flag waving), Discretending the Opponent (such as appeal to fear), Loaded language, and Appeal to Authority. We noticed that bias and propaganda overlap in their context as bias seen as systematic favoritism in presenting information or news and propaganda is to delibrately persuade audience to manupliate their opinion (Rodrigo-Ginés et al., 2023), we considered every biased post as propaganda. Meaning, in the context of social media posts and media in general, any biased opinion overlaps with the categories of propaganda and thus can be considered as propagandistic. However, if a post is unbiased, it still can hold propagandistic meaning. Following are the main categoreis of propaganda:

1. Appeal to authority. A claim is considered true (i.e., propagandistic) if a valid authority or expert on the issue said it was true.

2. Appeal to fear/prejudices. Supporting an idea by instilling anxiety and/or panic in the population towards an alternative.

3. Exaggeration/minimization. Making things larger, smaller, better or worse than what it really is.

4. Flag-waving. Playing on strong national feeling (or to any group, ethnicity, gender, race, religion, or political preference) to justify an action, reaction or an idea.

5. Virtue. Words or symbols in the value of target audience that produce a positive image when attached to a person or issue. Example words such as safety, peace, hope, happiness, security, leadership, freedom, "*The Truth*" are some virtue words.

6. Loaded language. Using phrases/words with strong emotional implications (can be positive or negative) to influence the audience.

7. Slogans. Striking phrases, typically short, that may include labeling and stereotyping. Slogans tend to act as emotional appeals.

8. Repetition. Repeating the same message multiple times will make the audience to accept it eventually.

9. Reductio ad hitlerum. Persuading an audience to disapprove an idea or an action by suggesting that the idea is popular among groups hated in contempt by the target audience.

10. Red Herring (presenting irrelevant data). Introducing irrelevant material to the issue being discussed, so that the attention of audience is diverted away from the points made.

11. Labeling. Labeling the object of the propaganda campaign as something that the target audience hates, fears, finds undesirable or, contrarily, loves, praises.

12. Black-and-white fallacy or dictatorship. Presenting two alternative options as the only available possibilities, when in fact more possibilities exist.

2.2 Data Annotation Process

FIGNEWS 2024 shared task provides with a shared spreadsheet including 15000 posts, and categorical entries for bias and propaganda labels. The category labels for bias subtask are:

1. Unbiased: In accordance with the guidelines mentioned in the previous section, the posts are evaluated according to the intentions of the author or posts. For example if a post shares news report from media outlets or if it shares a statement from some officials without presenting subjective opinion of the author, then its labeled as unbiased even if the statement itself might be biased.
2. Biased against Palestine.
3. Biased against Israel.
4. Biased against both Palestine and Israel.
5. Biased against others.
6. Unclear: When the text is too short or if the post includes a video that cannot be viewed by the annotation team.
7. Not Applicable: When the post it does not imply a meaning related to the Israel/Gaza war.

The category labels for propaganda subtask:

1. Propaganda: Some posts were labeled as propaganda although they are not labeled as biased, this is because people sometimes circulate propaganda unknowingly. For example, when

someone reports a speech for general or military commander, they repost the speech without doing any editing and without presenting their personal views on the matter. If the speech itself is propaganda, we decided to label such posts as propaganda without labeling it as biased.

2. Not Propaganda.

3. Unclear: When the text is too short or if the post includes a video that cannot be viewed by the annotation team.

4. Not Applicable: When the post is unrelated to the Israel/Gaza war.

The procedure followed in annotating posts is summarized as follows:

- A Team member logs into the files, scrolls down to reach the last annotated line, then writes his/her ID number at the proper cell, then reads the post in preferred language.
- The first thing the team tries to recognize is the use of hate speech, name calling and labeling, if bias is identified, the annotator would select the proper label.
- If the no such phrases found, the annotator tries to recognize double standards in the post
- The annotator pays attention to cases of quotes and tries to distinguish between posts presenting the author's opinion or if they are circulating information from biased sources.
- In some cases where there is too much ambiguity, the annotator would leave the post to be discussed with the annotating team
- The team tries to ensure that 10% of the annotated posts are in Inter-Annotator Agreement.

2.3 Inter-Annotator Agreement Analysis

FIGNEWS 2024 has split the 15000 posts into 15 batches, and it dedicates 10% of posts for Inter-Annotator Agreements analysis (IAA). The dedicated posts for annotators are duplicate and each annotator performs the labeling individually. The aim of IAA is to measure the efficiency of the guideline and how did they annotators follow it to annotate the posts.

We applied the Cohen's Kappa IAA measure (Cohen, 1960). The values of the agreements among the team's annotators for annotating posts as biased/non-biased are shown in table 1and for annotating propaganda/non-propaganda are shown in table 2.

	A1	A2	A3	A4	A5
A1		0.48	0.47	0.55	0.54
A2	0.48		0.46	0.56	0.39
A3	0.47	0.46		0.50	0.44
A4	0.55	0.56	0.50		0.43
A5	0.54	0.39	0.44	0.43	

Table 1: The Cohen's Kappa IAA scores for the 5 annotators in the team for the Bias annotation task.

	A1	A2	A3	A4	A5
A1		0.26	0.48	0.48	0.38
A2	0.26		0.22	0.23	0.15
A3	0.48	0.22		0.48	0.34
A4	0.48	0.23	0.48		0.31
A5	0.38	0.15	0.34	0.31	

Table 2: The Cohen's Kappa IAA scores for the 5 annotators in the team for the Propaganda annotation task.

From the tables above, it can be seen that for annotating bias, the scores show a moderate agreement between annotators. Landis and Koch interpreted the outcomes of the Cohen's Kappa measure as follows: a score less than zero indicates that there no agreement; 0–0.2 indicates a slight agreement; 0.21–0.4 is a fair agreement, 0.41–0.6 is moderate; 0.61–0.8 is substantial agreement, and if the score is greater than 0.81 is considered substantial to 1, which shows a perfect agreement (Landis & Gary, 1977).

As for the task of annotating posts into propaganda/non-propaganda, the scores of Cohen's Kappa range between 0.15 to 0.48, indicating an agreement ranges between slight (in only one score), fair, and moderate. It can be noticed that the bias posts could be identified in an easier way since they are easier to interpret and one can correlate realistically with the news. Propaganda, on the other hand, has a larger number of categories to be classified to, and this may have process to annotation and categorizing posts into between annotators more challenging.

3 Team Composition and Training

Our team is composed of five staff members from the Department of Computer Science, Birzeit University. All members of the team are Palestinians from West Bank. Four of the five members are male and one female.

Since the team members are from the same department, it was convenient for us to hold

regular meetings and discussions, feedback, and trial and error sessions. The team leader has individually met with each team member and worked together for several hours discussing as many cases as possible and observing the adherence to mutual standards. The team would leave ambiguous posts to be discussed as a group.

It is worth to mention that some of the team members are specialists in NLP and ML, and since all members are Palestinians, they are deeply immersed in political dialogues concerning the Palestinian/Israeli conflict and they have been observing related news throughout their whole lives.

4 Task Participdsxaation, Results and Discussion

In this section, a comparison of corpora sizes is done between our teams' annotation output and related NLP and manual annotated corpora. The volume of the corpora plays a significant role in deciding the accuracy of the NLP models developed after training on the corpora. Therefore, in order to evaluate the output of our team, we have taken a look at the size of corpora used to train existing models, in order to develop a perception of how large should the corpus be in order to produce an impactful corpus.

The CyberEquity Team has managed to annotate ~3000 posts out of 15000, 10% of annotated posts are part of IAA set. Comparing our performance with related participations in such shared tasks, for example, Sora *et. al.* (Sora Lim, 2020) has covered 966 news articles, Çağrı (Çöltekin, 2020) user a large 36,000 records to detect offensive language in Turkish, but the team has manually annotated a small portion of the corpus. Marta *et. al.* (Marta Sabou, 2014) presents a similar shared task, in which they crowd sourced a best-practice guideline. However, the paper does not explicitly mention the size of the corpus. Hutto *et. al.* (Hutto, 2015) has used an average of 91 human annotated labels as test set.

Having looked at related work and having compared the volume of the corpora, it comes clear that 3000 records of manually annotated labels is an acceptable amount of data to train highly accurate NLP models. Moreover, our team has achieved competitive results compared with other participating teams in FIGNEWS 2024 Shared Task, especially in the propaganda subtask where our team has secured third rank out of sixteen teams in:

- Quantity of data points
- Quality of results measured by IAA agreement
- Consistency score measured by cross-team Macro F1 average.

5 Conclusion

The paper presents the process of manual annotation of Facebook posts written in five languages: Arabic, Hebrew, English, French, and Hindi. The labels adopted highlight the bias and propaganda in the coverage of Gaza-Israel war. The paper has proposed a comprehensive guideline to identify and label posts into bias and propaganda. The research work has been dedicated as a response to FIGNEWS 2024 call for participation in two subtasks. Our team has participated in annotating 3000 different posts, including 10% for IAA agreements analysis. Comparing the volume of our annotated data with related earlier work on both manual and computational annotation, it can be concluded that the volume produced can be used to train ML models and can be used to evaluate clustering models.

6 References

Çöltekin, Ç. (2020). A Corpus of Turkish Offensive Language on Social Media. *the Twelfth Language Resources and Evaluation Conference* (pp. 6174–6184). Marseille, France: European Language Resources Association.

Alam, F., Mubarak, H., Zaghouani, W., Da San Martino, G., & Nakov, P. (2022). Overview of the WANLP 2022 shared task on propaganda detection in Arabic. *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, (pp. Abu Dhabi, UAE).

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20(1)*, 37-46.

D. Abuaiadah, D. R. (2017). Clustering Arabic Tweets for Sentiment Analysis. *IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, (pp. 449-456). Hammamet, Tunisia.

Da San Martino, G., Seunghak, Y., Barrón-Cedeno, A., Barrón-Cedeno, R., Petrov , R., & Preslav , N. (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)* (pp. 5636-5646). Association for Computational Linguistics.

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., . . . Giovanni, D. (2021). SemEval-2021 task 6: Detection of persuasion techniques in texts and images. *Proceedings of the15th International Workshop on Semantic Evaluation, SemEval '21*, (pp. 6603–6617). Bangkok, Thailand.

FAIR.org. (n.d.). Retrieved 2024, from https://fair.org/take-action-now/media-activism-kit/how-to-detect-bias-in-news-media/

Hamborg, F. D. (2019). Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 391-415.

Hutto, C. F. (2015). Computationally detecting and quantifying the degree of bias in sentence-level text of news stories. *Second International Conference on Human and Social Analytics.* Barcelona, Spain.

Landis, J., & Gary, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Marta Sabou, K. B. (2014). Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. *the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 859–866). Reykjavik, Iceland: European Language Resources Association (ELRA).

Park, S. a. (2009). NewsCube: delivering multiple aspects of news to mitigate media bias. *In Proceedings of SIGCHI on Human Factors in Computing Systems* (pp. 443–452). New York, NY, USA: Association for Computing Machinery.

Rodrigo-Ginés, F.-J., Carrillo-de-Albornoz, J., & Plaza, L. (2023). Hierarchical Modeling for Propaganda Detection: Leveraging Media Bias and Propaganda Detection Datasets. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023).* Jaén, Spain.

Shahi, G. M. (2022). AMUSED: An Annotation Framework of Multimodal Social Media Data. *Intelligent Technologies and Applications* (pp. 287-299). Springer International Publishing.

Sora Lim, A. J. (2020). Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. *the Twelfth Language Resources and Evaluation Conference* (pp. 1478–1484). Marseille, France: European Language Resources Association.

Williams, A. (1975). Unbiased Study of Television News Bias. *Journal of Communication*, 190-199.

Zaghouani, W., Jarrar, M., Habash, N., Bouamor, H., Zitouni, I., Diab, M., . . . AbuOdeh, M. R. (2024). The FIGNEWS Shared Task on News Media Narratives. *Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).* Bangkok, Thailand: Association for Computational Linguistics.