The CyberEquity Lab Annotation Team:

Our Classification Guidelines

FIGNEWS-2024

The CyberEquity Lab is an annotation team consisting of five staff members of the Computer Science Department, Birzeit University, Palestine. Normally, the affiliation and nationality of a group of academics working on guidelines for bias and propaganda classification/annotation are irrelevant, however in our case, it is important to point out that we are a Palestinian team classifying the bias and propaganda in Facebook posts covering the Israel-Gaza war, that's because maintaining objectivity in such situation where we are impartial part of Palestinian society could be pretty challenging. Addressing this challenge, our team has worked hard to lay clear guidelines and to study the literature to find objective definitions for the words bias and propaganda, and how to identify bias and propaganda in particular speech.

Proposed Guidelines:

Proposed Guidelines 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 when trying to justify
2. Hate Speech: Oxford Languages' definition is "abusive or threatening speech or writing that expresses prejudice on the basis of ethnicity, religion, sexual orientation, or similar grounds."
3. Double Standards: holding two different groups onto different sets of standards which implies favoritism and bias towards one side over the other.
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 therefore dehumanize them.

Proposed Guidelines for Classifying Propaganda:

1. Appeal to authority: A claim is considered true(i.e., Propaganda) because 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 / minimisation: Making things larger, better, worse than what it really is.
4. Flag-waving: Playing on strong national feeling (or to any group, such as gender, race, religion, or political preference) to justify an action or an idea.
5. Virtue: Words or symbols in the value of the target audience that produce a positive image when attached to a person or issue. Swords such as Peace, hope, happiness, security, wise leadership, freedom, “The Truth”, etc. are virtue words.
6. Loaded language: Using phrases/words with strong emotional implications (either positive or negative) to influence the audience.
7. Slogans: Short striking phrases that may include labeling and stereotyping. Slogans tend to act as emotional appeals.
8. Repetition: Repeating the same message over and over again, so that the audience will eventually accept it.
9. Reductio ad hitlerum: Persuading an audience to disapprove an action or an idea by suggesting that the idea is popular with groups **hated** in contempt by the target audience.
10. Red Herring (presenting irrelevant data): Introducing irrelevant material to the issue being discussed, so that everyone’s attention is diverted away from the points made.
11. Labeling: Labeling the object of the propaganda campaign as something that the target audience fears, hates, finds undesirable or loves, praises.
12. Black-and-white fallacy or dictatorship: Presenting two alternative options as the only possibilities, when in fact more possibilities exist.

Annotation Category Labels:

Subtask 1: Bias category labels

1. Unbiased: in accordance with the guidelines mentioned in the previous section.

   The posts are evaluated according to the intentions of the author, for example if a post shares news reports from media outlets or if it shares a statement from some official without presenting the author's subjective opinion, then it's labeled as unbiased even if the statement itself might be biased.
2. Biased against Palestine: Plenty of the annotated posts dehumanize Palestinians, or ignore the suffering of the thousands of Palestinian innocents while setting the scene on suffering from Israel’s side. Covering the suffering of Israeli civilians is certainly not by itself a bias, but when a post sympathizes with an Israeli victim from October 7th while ignoring the suffering of the hundreds of thousands of Palestinians, that is a form of bias.

   The posts that condemn one side and ignore the other for doing the same practice are considered biased, for example the targeting of civilians, posts often condemn it when done by one side and ignore it when done by the other.

   Another example: the use of phrase “terrorist”: terror means attacking civilians intentionally as a tactic of war, its proven from this conflict that both sides has targeted

civilians, however, plenty of post single Hamas out as terrorist and keep a blind eye on IDF's deliberate destruction of thousands of Palestinian lives, such posts are labeled biased for following double standards.
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 is unrelated to the Israel-Gaza conflict.

Subtask 2: Propaganda category labels

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, now if the speech itself is propaganda, our team has 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 conflict.

The process

- 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 prefere 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.

Training and Support

The team leader has individually met with each team member and worked together for a couple of 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.

Ethical Consideration

As stated earlier, our team is made up of Palestinian academics, which makes it challenging to maintain objectivity when we are not impartial, and more importantly, we need to be wary of uncritically using our own experience and perception of the conflict when classifying posts covering the conflict. For example, we might read posts circulating proven misinformation, or in other cases, circulating opinions that Palestinians reject. Our team has kept constant dialogues and discussions to point out the exact narratives that can be objectively deemed as misinformation and propaganda.

Ensuring Consistency

Consistency between the teammates and also throughout the whole data has been maintained by clearly defining the labels and criteria, and by teamwork. Our team has made sure to meet frequently enough and consult each other when there is ambiguity. The team members would often skip ambiguous posts to discuss them together later. We have highlighted many different examples and agreed to a unified standard.