Annotation Guidelines: Fallacious Arguments Classification in Twitter (X)

1 Introduction

The purpose of this annotation task is to identify fallacies present in tweets [1]. Annotators will be provided with a tweet and some context to determine if the tweet contains any fallacies from a predefined list.

Fallacies Standard dictionaries define fallacies as misleading arguments or defective reasoning [Oxford English Dictionary, 2023]. Modern approaches of argumentation theory state that fallacies are "deficient moves in argumentative discourse" [van Eemeren, 2001]. Such faulty reasonings are divided into plenty of categories and subcategories. We shortlisted some fallacy types for our annotation process based on their prevalence in our dataset and previous studies related to fallacy detection and classification.

Dataset The dataset contains tweets in English associated to the COVID pandemic and politically related topics. Twitter's API was used to extract tweets from the 25th of March 2020 until the 25th of March 2021. The keywords for the extraction were: covid, azithromycin, invermectin, bleach, vaccin, moderna, astrazeneca/astra zeneca, 5g, artificial, lab made, lab created, laborator made, laborator weapon, biological weapon, blood type, pandemic, army, armies, militar mobiliz, militar mobilis, h1n1. Data transformation and cleaning processes were applied to remove duplicates and redundancies, anonymize user information, retrieve the long-text version of each tweet, among others. Topic modeling techniques were used to filter tweets that deviate from the main topics of interest.

2 Context Information

Annotators will be presented with a *main tweet* and up to six tweets that preceded or followed the main tweet, which we call *context tweets*. That is, the *main tweet* was a reply or a quote of the preceding *context tweets*, or vice-versa. The context tweets are provided to offer a better understanding of the conversation in which the main tweet took place. Below is an example of how context information will be presented. Notice that the *main tweet* is highlighted to aid the annotator.

The anonymised user information is also available. This helps to identify, for instance, if the tweets correspond to a conversation between users, or a thread. Also to preserve the anonymity of users, we replace the mentions of users in the tweet's text with the placeholder

[1] When the data was extracted the platform was still called Twitter.

"@user". It is usual for tweets to contain many consecutive user mentions. In this cases, if more than three user mentions appear consecutively, we replaced by "@user @user ... @user"

Example 1 (a conversation):

> [context] [user1]: @user Deranged? Like this statement was sane? "And then I saw the disinfectant, where knocks it out in one minute, and is there a way we could do something like that by injection inside or almost a cleaning. ...so it would be interesting to check that."
>
> [context] [user2]: @user Remember the comment about moderately intelligent people seeing the context of what's said? Dems, Trump told us to inject disinfectant... No he didn't, don't be dull and spin it to suit your narrative. Liberals are irrational, they cannot control their emotions and need help
>
> **[main tweet] [user1]: @user How are you able to TRY and dispute video proof? HE SAID IT! This is the "poorly educated" thing that I refer to.**
>
> [context] [user2]: @user He didn't tell anyone to do it, he was talking about testing it. Dem supporters came out and said he told people to inject disinfectant, he didn't stop spinning bullshit. The moderately intelligent can see the context
>
> [context] [user1]: @user What kind of a fool would even consider testing the injection of disinfectants? Let alone say that it might be interesting to try it. Ignorance abounds in the Chump cult.

Example 2 (a thread):

> [context] [user123]: @user @user ... @user There are several #coronavirus vaccines for animals. While a vaccine for #SARSCoV2 is not a given, two encouraging data offer hope for #COVID19 Oxford MERS vaccine w/ same ChAdOx1 vector "Old-fashioned" inactivated vaccine in NHP
>
> **[main tweet] user123: 1/Coronavirus Tweets from the Experts, 4/26/2020**
>
> [context] [user123]: 2/A small bit of good news
>
> [context] [user123]: 3/Antivaxxers: Still probably the worst people in America.

3 Fallacy identification

Annotators should decide whether the main tweet contains one or more of the following fallacies: loaded language, appeal to fear, appeal to ridicule, hasty generalization, ad hominem, and false dilemma. **Do not consider fallacies in the context tweets; focus solely on the main tweet.** If more than one fallacy is present in the tweet, select all the types you consider the tweet contains. If none of the previous fallacies is present, the annotator should select "None of the above."

4 Types of Fallacies

Below, we list the fallacy types that should be considered for this annotation process. Note that subclasses for some categories are not exhaustive.

1. **Loaded Language:** The use of words and phrases with strong connotations (either positive or negative) to influence an audience and invoke an emotional response [Weston, 2018, Goffredo et al., 2022].

   Example:

   (a) It was a stupid musing about what doctors or researchers might explore. It's just idiotic to think he meant for people to go out and buy hypodermics and inject themselves. Is that even possible?

   (b) He will hijack this boycott. He would hijack anything that gives him attention. That's a given.

   (c) I could go weeks while Obama was POTUS without barely hearing his name as he took care of business. Now it is like an ever present rolling chyron and constant iphone alerts with Trump's dumpster fire.

   In example 1.a, the user uses the words "stupid" and "idiotic", which carry strong negative connotations. Likewise, in example 1.b, the word "hijack" implies that the subject of the tweet is opportunistic and manipulative. Using this word to describe someone's actions suggests that their involvement is not just negative, but actively malicious and aggressive. Example 1.c uses the phrases "dumpster fire" and "ever present rolling chyron and constant iPhone alerts" to imply chaos and a sense of overwhelming and pervasive presence respectively.

2. **Appeal to Fear (argumentum in terrorem):** Eliciting fear to support a claim [Walton, 1987, Goffredo et al., 2022].

   Examples:

   (a) Meanwhile, the Israelis, pulling out their hair over coronavirus, and restaurants even smashing plates, have started car-bombing each other...

   (b) The truth will come out, there is censorship on medics and science, ivermectin and hydroxychloroquine with Zinc and vitamins have been proven to be a preventative against covid, but if you're happy to take the vax then fine, just hope you're still here to see the truth.

   (c) Derek Rossi of Harvard Founder of Moderna, who never made meds & will probably get the #COVID19 vaccine contract invested in by #BillGates & Fauci, has made it possible to genetically modify us altering our genome changing the function of our stem cells.

   In example 2.b, the user aims to manipulate the reader's emotions by implying that there is a suppression of information and suggesting dire consequences for those who choose to receive the COVID-19 vaccine. Similarly, the argument in 2.c relies on fear and alarmist claims rather than factual evidence. It does not provide any concrete evidence to support the claim that the COVID-19 vaccine will genetically modify humans or alter our genome.

3. **Appeal to Riducule (ad absurdo):** Presenting an opponent's argument as absurd, ridiculous, or humorous. Mocking the opponent's point of view. A common instance is to use the expression "That's crazy!" to dismiss an argument [Bock, 2018].

Notice that although “That’s crazy!” is an instance of appeal to ridicule, “You are crazy!” is an instance of ad hominem.

Examples:

(a) The COVID guidelines be like “Make sure you touch a coffee cup with three fingertips when lowering your mask or the virus will mutate.” It’s OCD and witchcraft kind of stuff at this point.

(b) user1: The Biden-Harris administration will get COVID-19 under control by listening to the experts, implementing nationwide testing and tracing, and ensuring vaccines are safe and free for all.
user2: You mean the plan that’s already in place?

(c) If you still believe that The Chinese government didn’t make the #Covid_19 #ChinaVirus you should consider drinking bleach while eating a cactus

The tweet in Example 3.a attempts to dismiss a precautionary measure regarding COVID-19 hygiene protocols by ridiculing it as “OCD and witchcraft.” It exaggerates and caricatures the recommended hygiene practices, presenting them in a manner that makes them seem absurd. In Example 3.b user2 uses sarcasm to mock the proposed plan for controlling COVID-19 by making it sound redundant or obvious. In Example 3.c the speaker aims to make the belief (that the Chinese government did not fabricate COVID) seem unworthy of serious consideration by associating it with an absurd and harmful action (drinking bleach and eating a cactus).

4. **Hasty Generalization (overgeneralizing):** Making a broad statement about a group or population based on a limited or unrepresentative sample. It usually follows the form: X is true for A, X is also true for B, therefore, X is true for C, D and E [Weston, 2018, Sahai et al., 2021].

Examples:

(a) Liberals are irrational, they cannot control their emotions and need help.

(b) Ivermectin KILLS BAD #COVID-19 IN 2-6 DAYS: my 90-yro Aunt, on edge of intubation, ICU, got rid of it in 5 days; feeling better after 1. Get it approved!

(c) I’m seeing more and more reporting on Covid vaccines being wasted because of Govt “rules”; this is exactly why Govt should not be allowed near anything outside of upholding/enforcing the Constitution-especially healthcare.

Example 4.a makes a broad statement about liberals without providing enought evidence to support that the whole group is irrational. In example 4.b, the user makes a generalization about the effectiveness of ivermectin as a treatment for COVID-19 based on a single anecdote. Similarly, 4.c makes a generalization about the government’s ability to manage healthcare based on a single report of COVID-19 vaccine wastage.

5. **Ad Hominem (Ad Personam):** Attacking the person or some aspect of the person making the argument rather than addressing the argument itself [Habernal et al., 2018, Weston, 2018, Walton, 1987]. We consider the subcategories listed below. If the tweet contains any of the subcategories, the annotator should select *Ad hominem.*

a. **Abusive Ad Hominem:** A pure attack on the character of the opponent. It often takes the form of “My opponent is not a good person or has a bad trait, therefore his or her argument should not be accepted” [Walton, 1998, Habernal et al., 2018, Walton, 1987].

Example:

> Keep proving that you’re an uneducated deplorable by showing your lack of reading comprehension.

b. **Tu Quoque Ad Hominem:** Suggesting that because someone does not practice what they preach, their argument is invalid. It’s analogous to a “You did it first” “You are just the same” or “You are just as bad” kind of attack [Walton, 1998, Habernal et al., 2018, Goffredo et al., 2022].

Example:

> You can’t bring up someone else’s whataboutism if/when you use it yourself. That would be called HYPOCRISY!

c. **Bias Ad Hominem:** Implying that the opponent is personally benefiting from his stance in the argument. That is, questioning the impartiality of the arguer [Walton, 1998, Habernal et al., 2018, Goffredo et al., 2022].

Example:

> Follow the money. Rishi Sunak refuses to say if he will profit from Moderna Covid vaccine.

d. **Name-calling, Labelling:** Using derogatory language or offensive labels against the opponent to undermine an argument instead of addressing the content of the argument [Goffredo et al., 2022].

Example:

> What kind of a fool would even consider testing the injection of disinfectants? Let alone say that it might be interesting to try it. Ignorance abounds in the Chump cult.

Note that “Chump” is a derogatory term to refer to Donald Trump that takes advantage of the resemblance of his last name to the word chump (a foolish or easily deceived person).

6. **False Dilemma (Black and White Fallacy, Bifurcation Fallacy):** Presenting a situation as having only two alternatives, when in reality there are more options available. It oversimplifies a complex issue by reducing it to only two possible outcomes or choices, often in a way that excludes other possibilities, nuances, or middle-ground [Sahai et al., 2021, Culver, 2018, Da San Martino et al., 2020].

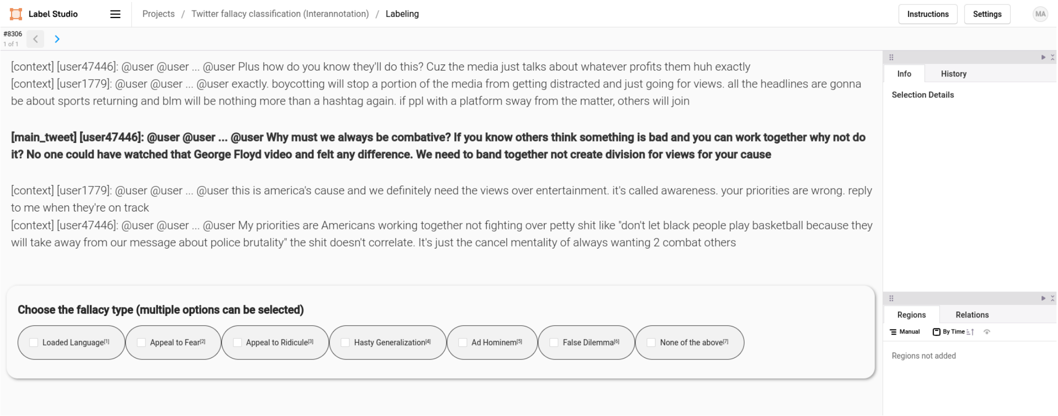
Examples:

> (a) Don’t let people die in hospitals from COVID-19 when #ivermectin is available.
>
> (b) How are you helping small business & mothers handle Covid crisis if you’re helping people cheat & skirt the system meant to slow the virus spread so small business can get back to business & mom’s can get kids back to life? Seems you’re the danger.

Example 6.a suggests that there are only two options: let people die from COVID-19 in hospitals or use ivermectin as a treatment. This oversimplification ignores other potential treatments, medical guidelines, and the complexities of medical decision-making. Example 6.b presents a binary choice between strictly adhering to COVID-19 guidelines or being labeled as someone who hinders small businesses and mothers.

5 Annotation tool

The annotation process takes place in the Label Studio platform. Label Studio is a free data labeling tool. Figure 1 shows an instance of how the tweet's information is displayed and the multiple-choice buttons that the annotator can use.



[context] [user47446]: @user @user ... @user Plus how do you know they'll do this? Cuz the media just talks about whatever profits them huh exactly
[context] [user1779]: @user @user ... @user exactly. boycotting will stop a portion of the media from getting distracted and just going for views. all the headlines are gonna be about sports returning and blm will be nothing more than a hashtag again. if ppl with a platform sway from the matter, others will join

[main_tweet] [user47446]: @user @user ... @user Why must we always be combative? If you know others think something is bad and you can work together why not do it? No one could have watched that George Floyd video and felt any difference. We need to band together not create division for views for your cause

[context] [user1779]: @user @user ... @user this is america's cause and we definitely need the views over entertainment. it's called awareness. your priorities are wrong. reply to me when they're on track
[context] [user47446]: @user @user ... @user My priorities are Americans working together not fighting over petty shit like "don't let black people play basketball because they will take away from our message about police brutality" the shit doesn't correlate. It's just the cancel mentality of always wanting 2 combat others

Choose the fallacy type (multiple options can be selected)

□ Loaded Language[1] □ Appeal to Fear[2] □ Appeal to Ridicule[3] □ Hasty Generalization[4] □ Ad Hominem[5] □ False Dilemma[6] □ None of the above[7]

Figure 1: Example of Label Studio's annotation interface.

References

Oxford English Dictionary, December 2023. URL `https://doi.org/10.1093/OED/5071256414`.

Frans H van Eemeren. *Crucial concepts in argumentation theory*. Amsterdam University Press, 2001.

Anthony Weston. *A Rulebook for Arguments*. Hackett Publishing, February 2018. ISBN 978-1-62466-655-1. Google-Books-ID: XhVNDwAAQBAJ.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. Fallacious Argument Classification in Political Debates. In *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, pages 4143–4149, Vienna, Austria, July 2022. International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2022/575. URL `https://hal.science/hal-03873412`.

Douglas N. Walton. *Informal Fallacies*. John Benjamins Publishing, January 1987. ISBN 978-90-272-7890-6. Google-Books-ID: LQVCAAAAQBAJ.

Gregory L. Bock. Appeal to Ridicule. In *Bad Arguments*, pages 118–120. John Wiley & Sons, Ltd, 2018. ISBN 978-1-119-16581-1. doi: 10.1002/9781119165811.ch18. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119165811.ch18`. Section: 18 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119165811.ch18.

Saumya Sahai, Oana Balalau, and Roxana Horincar. Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 644–657, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.53. URL `https://aclanthology.org/2021.acl-long.53`.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 386–396, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1036. URL `https://aclanthology.org/N18-1036`.

Douglas Walton. *Ad Hominem Arguments.* University of Alabama Press, September 1998. ISBN 978-0-8173-0922-0.

Jennifer Culver. False Dilemma. In *Bad Arguments*, pages 346–347. John Wiley & Sons, Ltd, 2018. ISBN 978-1-119-16581-1. doi: 10.1002/9781119165811.ch81. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119165811.ch81`. Section: 81 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119165811.ch81.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL `https://aclanthology.org/2020.semeval-1.186`.