Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions

Saumya Yashmohini Sahai, Oana Balalau, Roxana Horincar

Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions

Saumya Yashmohini Sahai
The Ohio State University, USA
sahai.17@osu.edu

Oana Balalau
Inria, Institut Polytechnique de Paris, France
oana.balalau@inria.fr

Roxana Horincar
Thales Research & Technology, France
roxana.horincar@thalesgroup.com

Abstract

People debate on a variety of topics on online platforms such as Reddit, or Facebook. Debates can be lengthy, with users exchanging a wealth of information and opinions. However, conversations do not always go smoothly, and users sometimes engage in unsound argumentation techniques to prove a claim. These techniques are called fallacies. Fallacies are persuasive arguments that provide insufficient or incorrect evidence to support the claim. In this paper, we study the most frequent fallacies on Reddit, and we present them using the pragma-dialectical theory of argumentation. We construct a new annotated dataset of fallacies, using user comments containing fallacy mentions as noisy labels, and cleaning the data via crowdsourcing. Finally, we study the task of classifying fallacies using neural models. We find that generally the models perform better in the presence of conversational context.We have released the data and the code at github.com/sahaisaumya/informal_fallacies.

1 Introduction

Argumentation plays a critical part in our lives as it helps us make decisions and reason about the world around us. Studies (Sanders et al., 1994) have shown that learning how to argue increases the ability to identify weak arguments and decreases the tendency to use verbal aggressiveness. Fallacies are weak arguments that seem convincing, however, their evidence does not prove or disprove the argument's conclusion. Fallacies are usually divided into formal and informal, where the former can be easily described using logical representations, while for the latter, an analysis of the content is more appropriate. Fallacies are prevalent in public discourse. For example, The New York Times labeled the tweets of Donald Trump between 2015 and 2020 and found thousands of insults addressed to his adversaries. If made in an argument, an insult is an ad hominem fallacy: an attack on the opponent rather than on their argument. In private conversations, other types of fallacies might be more prevalent, for example, appeal to tradition or appeal to nature. Appeal to tradition dismisses calls to improve gender equality by stating that "women have always occupied this place in society". Appeal to nature is often used to ignore calls to be inclusive of the LGBTQ+ community by stating "gender is binary". The underlying premises of such arguments are "traditions are correct" and "what occurs in nature is good".

Creating a dataset of fallacious arguments is difficult, given that there are over 100 types of fallacious arguments (Scalambrino, 2018). There have been several attempts to create comprehensive datasets: Habernal et al. (2017) proposed a game in which players add fallacies in the hope of fouling other participants, in Habernal et al. (2018a) ad hominem fallacies are found using a subreddit's rule violations, while in Da San Martino et al. (2019) fallacies are annotated together with other propaganda techniques in news articles. However, our work is the first to propose a viable solution for finding fallacious arguments belonging to many different fallacy types.

In this work, we study fallacies in public discussions on online forums. Our salient contributions are: *i*) we align informal fallacies mentioned on Reddit within the pragma-dialectic theory of argumentation (van Eemeren and Grootendorst, 1995); *ii*) we design a methodology for mining and labeling easily fallacies in online discussions; *iii*) we construct a large and balanced dataset of fallacious arguments; *iv*) finally, we evaluate several neural

Part of this work was done while the first author was an intern at Inria, France.

models on the task of predicting fallacious arguments, and we find that taking into consideration additional conversational context is important for this task.

2 Background

2.1 Fallacies in Argumentation Theory

Humans use argumentation when they evaluate the validity of new ideas, or they want to solve a difference of opinion. An argument contains: *i*) a proposition called claim, conclusion or standpoint, to be validated; *ii*) the premises called also evidence, which are the backing propositions; *iii*) an inference relation between the evidence and conclusion that validates or disproves the conclusion. A fallacy is a flawed argument, where the inference relation or the premises are incorrect. Fallacies are generally divided into formal and informal fallacies. Formal fallacies are arguments that can be easily represented as invalid logical formulas, such as *denying the antecedent*, which is a wrong application of modus tollens. Although many informal fallacies can be also represented as invalid arguments, informal fallacies are easier to describe and understand without resorting to logical representations (Hansen, 2020).

In this work, we follow the **pragma dialectic theory of argumentation**. The theory developed by van Eemeren and Grootendorst (1995) views argumentation as a complex speech act. The dialectical aspect is represented by two parties who try to resolve a difference of opinion by engaging in a discussion, each party making a move towards resolution. The pragmatic aspect describes the moves in the discussion as speech acts, more precisely as the illocutionary acts introduced by Searle (1979). van Eemeren and Grootendorst (1995) also developed ten rules which should guide argumentative discussions. The goal of the rules is to further the understanding of the difference of opinions and to create a fruitful discussion. For example, a rule states that parties must not prevent each other from advancing standpoints or from casting doubt on standpoints, while a second rule asks that a party may defend a standpoint only by advancing argumentation relating to that standpoint. An argument that prevents the resolution and thus violates one of the rules is a **fallacy**. In our work, we align frequent fallacies on Reddit with these rules, with the goal of formalizing their definitions.

Another well-known model that considers fallacies is the argumentation scheme introduced by Douglas Walton (Walton, 2005). A scheme consists of a conclusion, a set of premises, and a set of critical questions. The critical questions should be answered in order to prove that the premises support the conclusion, hence the argument is not a fallacy. For example, the scheme for an argument from expert opinion (Walton, 2005) has the premises *E is an expert in domain D, E asserts that A is known to be true, A is within D* and the conclusion *therefore, A may plausibly be taken to be true.* Some critical questions for this scheme are: *i*) Trustworthiness: Is E personally reliable as a source? *ii*) Backup Evidence: Is E's assertion based on evidence? Argumentation schemes have two main drawbacks: first, for each new fallacy, a new scheme should exist or be defined; and second, in the context of labeling an existing argument, many of the critical questions might be unanswerable as none of the parties discussed them.

2.2 Related Work

An initial effort for creating an extensive dataset of fallacies was made in Habernal et al. (2017). The authors created a platform for educative games, where players learn how to become better debaters. New fallacies are added to the platform by players that try to earn points by fouling other participants with invalid arguments. A follow-up on this work (Habernal et al., 2018a) mentioned a dataset of only around 300 arguments created via the platform, thus showing the need of finding other methods for creating larger datasets of fallacies.

Ad hominem fallacies in conversations have been addressed in (Habernal et al., 2018b). The authors used the subreddit ChangeMyView, which is a forum for civilized discussions, "a place to post an opinion you accept may be flawed, in an effort to understand other perspectives on the issue". The dataset of fallacies consists of comments that were removed by the moderators as they violated the rule of not being rude or hostile, hence committing an ad hominem fallacy.

Fallacious arguments are often made in the dissemination of propaganda. In Da San Martino et al. (2019), the authors annotate journal articles with 18 propaganda techniques, out of which 12 techniques are fallacies. Although an important resource in the study of fallacies, their labelling method and dataset have a few drawbacks. First, the dataset is highly unbalanced with 6 fallacies having a fair

number of mentions: name-calling (1294), appeal to fear and prejudice (367), flag-waving (330), causal oversimplification (233), appeal to authority (169), black and white fallacy (134), and 6 fallacies having less than 100 mentions: whataboutism (76), reductio ad hitlerum (66), red herring (48), bandwagon (17), labeling, obfuscation or intentional vagueness (17), straw men (15). Second, the task of finding the correct label for a span of text from a large set of labels (18 in their case) is intellectually complex and time-consuming. Our work focuses on collecting and annotating a balanced dataset of fallacy mentions while providing a methodology that can easily scale to a larger number of fallacies. In our approach, an annotator has to just verify that a comment contains one type of fallacy. In addition, we target fallacies in online conversations, where the style of argumentation is less structured than in a journal article.

3 Fallacies on Reddit

Finding a large sample of fallacious arguments is a challenging task as it assumes going through long conversations, finding arguments, and then verifying if the arguments are sound. Another major issue, even if we recognize the argument is flawed, is to find the exact fallacy that is committed, given that more than 100 types of fallacies have been proposed in the literature (Scalambrino, 2018).

Our goal is to construct an annotated dataset of fallacies using a mixed strategy: *i*) first, as noisy labels, we leverage user comments that mention the name of a fallacy, and second, *ii*) we clean this dataset by removing false-positive samples via crowdsourcing. Our intuition is that a person will mention a fallacy as a reply to another comment to highlight that the previous comment's argument is fallacious, as shown in Figure 1. This might not always be the case, as users could discuss fallacies in general, hence the need to further label the discussion using crowdsourcing.

We use the Pushshift Reddit API (Baumgartner et al., 2020) to retrieve data from Reddit. The API allows searching comments and submissions by their IDs or by a set of keywords. We start by making an exhaustive list of fallacies informed by Wikipedia. We chose Wikipedia as a resource for creating the list of fallacies as it is one of the most well-known sources of information, hence a Reddit user could peruse it easily to understand what fallacy was committed in the discussion. For each

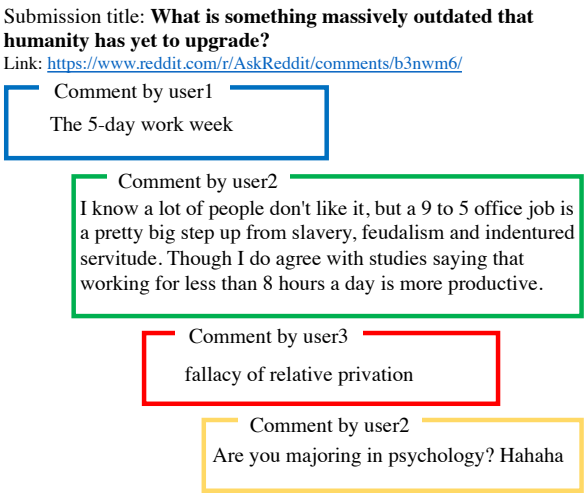

Figure 1: The redditor *user3* is pointing out a fallacy.

fallacy we find all its different designations, for example, appeal to tradition is also known under its Latin name, *argumentum ad antiquitatem*. We then do a keyword search for these fallacy types on Reddit comments, restricting the results to one year, May 2019 to May 2020. We retrieve in total $105K$ comments that match at least one fallacy. For comparison, in 2019, 1.7 billion comments were posted on Reddit. While it is very likely that many more posts contain fallacies, the small number of matches highlights the importance of choosing with care the comments to annotate. To understand in which subreddits people were more likely to mention names of fallacies, we compute the top 10 subreddits with the highest ratio of matched comments per number of subscribers, as shown in Table 1. The subreddits are broadly divided into subreddits on religion, morality, and science, with one subreddit dedicated to discussions on fallacies. The subreddits' focus is on debating, which involves creating, defending, and attacking arguments, therefore accusing the opponent of committing a fallacy might win you the debate.

From the list of most frequently mentioned fallacies we retained the top fallacies with more than 400 mentions, resulting in 32 fallacy types. This *shortlist of frequent fallacies* is presented in our **Appendix A**, with a definition, example, and argumentation rule violation (according to the pragma dialectic theory) for each fallacy. From this shortlist we do not consider the fallacies that were already studied in Habernal et al. (2018b), as their labeled dataset is also based on Reddit comments. We do not exclude fallacy types annotated in Da San Martino et al. (2019), as these are fallacious arguments

Subreddit	Description
Abortiondebate	A subreddit for debating abortion: ethics, religion, politics all welcome.
AskAChristian	A casual discussion forum - ask questions to Christians of various backgrounds.
fallacy	A subreddit on fallacies.
DebateVaccines	Debate and discuss issues surrounding vaccinations.
DebateEvolution	Reddit's premiere debate venue for the evolution versus creationism controversy.
Quraniyoon	Discuss the Qur'an Alone.
DebateReligion	A place to discuss and debate religion.
DebateAChristian	A curated community designed specifically for rational debates about Christian subjects.
AskConservatives	A subreddit for asking questions to conservatives.
DebateAVegan	A place for open discussion about veganism and vegan issues.

Table 1: Top 10 subreddits with highest ratio of comments mentioning a fallacy per number of subscribers.

in journal articles. We take random samples of 20 comments that mention one of our frequent fallacies and the comment to which they reply (the potential fallacious comment), and we check if the users have a good understanding of the respective fallacies. We keep the fallacies for which users generally had a correct sense of their definition. In addition, we filter fallacy types if more than 60% of potential fallacious comments were not true fallacious arguments. These conditions assure that the comments we will label have good quality and that we will find sufficient actual fallacy examples.

The remaining fallacies are selected for the creation of an annotated dataset of fallacies. These 8 fallacies are:

Appeal to authority / argument from authority fallacy / *argumentum ad verecundiam*. *Definition.* The claim is supported by the opinion of a person with authority, hence the claim is true. *Example.* Being vegan makes no sense because my father said so.

Appeal to majority / bandwagon argument / appeal to widespread belief / appeal to the people fallacy / *argumentum ad populum*. *Definition.* A claim is true because many people believe it to be true. *Example.* Being vegan makes no sense because so many of us are meat eaters.

Appeal to nature / naturalistic fallacy. *Definition.* An action A is justified/unjustified because it occurs/does not occur in nature. *Example.* Being vegan makes no sense as our body is designed for eating meat.

Appeal to tradition fallacy / *argumentum ad antiquitatem*. *Definition.* An action A is justified/unjustified because it has always been considered as such in the past. *Example.* Being vegan makes no sense as our ancestors have been meat eaters.

Appeal to worse problems / relative privation / not as bad as fallacy. *Definition.* There exists problem A that is worse than problem B, therefore B is justified.

Black-or-white / false dilemma / false dichotomy / bifurcation fallacy. *Definition.* In this argument, the claim is that only an event/action A should be considered. The first premise is that only two events, A and B are possible when there is at least a third event C possible. The second premise is that one of the events is bad, for example B, thus only event A should be considered. *Example.* You must wear a mask each time you go out, otherwise, you will die of COVID-19.

Hasty generalization fallacy. *Definition.* The claim is supported by insufficient evidence through inductive generalization. More precisely, we know that predicate P is true for a population sample, and we suppose it is true for the entire population. However, the sample is too small or it is not representative of the population. *Example.* The first week of September has been sunny, which means the rest of the month will be the same.

Slippery slope / thin edge of the wedge / camel's nose fallacy. *Definition.* A small event A will have a big unwanted consequence C. There is at least one more event B in the chain of causality (A will cause B, B will cause C), hence the slippery slope name of the fallacy. *Example.* If you break your diet and have one cookie tonight, you will just want to eat 10 cookies tomorrow and 20 the day after, and before you know it, you will have gained

back the 15 pounds you lost.

Rule violation. According to the pragma dialectic theory, an argument is a fallacy if it violates a critical discussion rule. The arguments above violate one of two rules, hence they are fallacies. The first rule violated states that defending a claim must occur through an appropriate argumentation scheme that is correctly applied. Argumentation schemes in van Eemeren and Grootendorst (1995) are different than schemes in Walton (2005). They are a formalization of the relation between the evidence presented and the standpoint to be defended. This rule is violated by all fallacies, except black-or-white. For example, in slippery slope, the argumentation is not valid as there is no clear causality chain between A and C. Black-or-white fallacy violates the rule that a party should not falsely present a premise as an accepted starting point, by stating that only events A and B are possible.

4 Dataset

Noisy labels. We used Amazon Mechanical Turk to create our annotated dataset. We selected 4 Master annotators, which had the highest agreement with the authors on identifying a set of fallacies (70 samples). An annotation task, defined as a HIT[1] consists of 10 items. Each item presents a sample extracted from a Reddit discussion. A Reddit discussion is started by a **submission**, e.g., a news article or a piece of text, to which users engage by writing **comments**. The comments and submission are organized in a tree-like structure: the submission is the root, and comments are nodes in the tree; we will use the terms grandparent, parent, and child to denote relations between comments. A sample given for annotation includes the **title** and the **link** of the original Reddit submission and **four comments**:

- the comment containing the mention of the fallacy (this is the **label comment**);
- the parent of the label comment, which should contain the fallacious argument (the **comment of interest or COI**);
- the parent of the COI, to give more context for the discussion;
- a direct reply to the label comment; preference was given to replies that had the same author as the COI; if no such comment existed, then we choose the top-rated comment.

[1]Human Intelligence Task on Amazon Turk

An example of a sample is shown in Figure 1.

For each fallacy described in Section 3, we retrieve all the label comments mentioning it and the context needed for creating a sample discussion (item). We keep the items for which: *i*) the comments are relatively short: the label comment has less than 500 characters (a shorter text will more likely be an accusation of committing a fallacy), and the other comments have less than 1000 characters; *ii*) we have enough context to understand the discussion: the COI is a direct reply to the submission or the child comment of a direct reply; *iii*) the COI or its parent do not contain the substring '*fallac*', a sign that this could be a discussion on fallacies and therefore the COI does not contain a fallacious argument, but it merely discusses or points out one. *iv*) we have access to the original discussion: the user or a moderator did not delete the comments, and the submission is not from a banned subreddit (the annotators can visit the link provided with the submission title); *v*) all the comments are in English.

Crowdsourcing task. Workers were presented with concise descriptions of the main concepts involved: argument, claim, evidence and fallacy. All the items in a HIT have to be annotated only for one fallacy. For example, we retrieved all the items where the label comment mentioned "hasty generalization fallacy" and we split them into HITs. We note that the fallacy committed in the comment might not be the same as the one signaled by the user. However, the authors have reviewed a large sample of comments (for the third vote explained further in this section) and did not encounter this situation. Hence, even if this might still occur, it should be rare. For each selected fallacy, we offered the definition together with an example of the fallacy, where we identified the claim and evidence. Furthermore, we instruct the workers not to label as a fallacy a comment that is sarcastic (sometimes accompanied by the explicit tag "/s") or a comment that is disproving the fallacy, e.g., *Who would think that we shouldn't become vegans just because our body is able to digest meat?*.

The workers are asked if the fallacy occurs in the comment of interest and if yes, they are prompted to highlight the corresponding text span. They are also asked to write the claim that is addressed by the comment of interest. Finally, they have to answer a question specific to each fallacy to prove their good understanding of the task. The ques-

tions are: *i*) appeal to authority: "What authority is being appealed to in the comment of interest, and hence is used as the basis for the argument?"; *ii*) appeal to majority: no question; *iii*) appeal to nature: "What natural phenomenon/event/activity is considered natural here?" *iv*) appeal to tradition: "What tradition is being appealed to in the comment of interest, and hence is used as the basis for the argument?"; *v*) appeal to worse problems: "Describe why the current problem (problem 1) is not a trivial issue." *vi*) black-or-white: "Name any additional alternative, which is possible but is not mentioned in the comment of interest." *vii*) hasty generalization: "Describe a case where the (hasty) generalization will fail." *viii*) slippery slope: "Please list any one event in the chain of slippery slope argument." By answering these questions, the workers would take the time to understand why the argument was a fallacy.

Annotated dataset. A HIT is annotated by two workers. We compute the Cohen's κ agreement for the task of deciding if a comment contains a fallacy (comment-level annotation), and γ inter-annotator agreement (Mathet et al., 2015) for the task of highlighting the tokens of the fallacy within the COI (token-level annotation), as shown in Table 2. For both measures, 1. implies perfect agreement. The comment-level annotation agreement varies from fair (black-or-white and hasty generalization) to substantial (appeal to authority), with the majority of fallacies in the moderate interval. The token-level agreement is moderate for appeal to worse problems and substantial for the rest.

Fallacy	**Comment** (Cohen's κ)	**Token** (γ)
Appeal to authority	0.64	0.68
Appeal to majority	0.47	0.79
Appeal to nature	0.60	0.74
Appeal to tradition	0.55	0.80
Appeal to worse problems	0.59	0.60
Black-or-white	0.40	0.68
Hasty generalization	0.38	0.71
Slippery slope	0.49	0.61

Table 2: Agreement between annotators.

In addition to the workers' votes, an expert annotator casts a third vote on comments, whenever there is a disagreement on the label. A comment is marked as fallacious if it has received two fallacy votes. The corresponding fallacious tokens of the comment are the union of the tokens highlighted by the annotators. We annotated comments until we reached roughly 200 fallacious comments per fallacy type. The details of the dataset are presented in Table 3.

Fallacy	**Number of comments**	**Mean tokens in spans**
Appeal to authority	212	21.49 ± 15.00
Appeal to majority	196	15.52 ± 11.55
Appeal to nature	208	15.16 ± 9.61
Appeal to tradition	210	16.35 ± 9.07
Appeal to worse problems	239	25.71 ± 17.44
Black-or-white	211	21.80 ± 14.77
Hasty generalization	204	19.76 ± 12.72
Slippery slope	228	27.98 ± 19.23
Overall	1708	20.69 ± 14.93

Table 3: Fallacious comments and tokens.

The total size of our **annotated dataset**, including comments and tokens that are non fallacious, consists of 3358 comments and $160K$ tokens. We observe that to find 1708 fallacious comments, we annotated only about two times more comments. This shows that our technique of finding fallacious comments is efficient.

We investigate if the label comment (i.e., the comment containing mention of the fallacy) is truly indicative of a fallacy in the COI. This can be useful for flagging the label comments that are likely to point to fallacious COI, therefore eliminating or reducing the need for crowdsourcing. Our intuition is that a classification method might differentiate when comments are accusations or just mention of fallacies. To investigate this, we used the fallacy/no-fallacy annotation as classes for label comment and trained a binary BERT classifier (Devlin et al., 2019). We obtained an F1 score of 67.41, indicating that the label comment's content is not sufficiently reliable. In conclusion, human annotators are still needed for annotating the true class of the COI.

Non fallacious comments. The comments for which two annotators confirmed they were not fallacious represent our annotated negatives (1650 comments). In order to have a more diverse set of negative examples, i.e. on similar and different topics, we construct a second set of negative examples (6400 comments) as follows. We retrieve all the users that wrote a label comment to a COI and the COI was identified as fallacious in the annota-

tion, our *gold users*. We take all their comments after the timestamp of the label comment that do not mention a fallacy name, and retrieve their parent comment. For each comment in the annotated dataset, we select one sample from our pool of parent comments from the same subreddit (if this exists) and one from a subreddit not seen in the annotated dataset. We retrieve a total of 6400 samples. These comments are used together with the annotated dataset, to create our **full dataset**, used to train classification models. The intuition of the sampling strategy is that, the gold users were able to recognize true fallacies at least one time, so they should spot other fallacies. Hence, if they reply to a comment without flagging it, the parent comment is likely to be non fallacious. There could be fallacious comments in this sample; however, we consider it less likely than a random sample.

5 Models and Discussion

Tasks. We address four tasks leveraging our annotated dataset, listed in the order of increasing granularity: *i*) **comment-level (CL) fallacy identification** (binary task of predicting if a comment is fallacious or not); *ii*) **comment-level fallacy type identification** (multi class prediction of the type of fallacy, with non-fallacious as one class in the 9 classes); *iii*) **token-level (TL) fallacy identification** (binary task of predicting if tokens in the COI belong to a fallacy or not); *iv*) **token-level fallacy type identification** (multi class prediction of tokens in the COI into one of the eight fallacy classes or the non-fallacy class).

5.1 Models

Random. We generate predictions by respecting the class distributions in the training set.

BERT. We fine-tune BERT by adding a linear layer on top of generated contextual representations. We use the token level embedding in token detection tasks and `[CLS]` embedding in the case of classification tasks.

MGN. We adopt the best architecture reported in Da San Martino et al. (2019), which is a multi-granularity network that uses lower granularity sentence-level (which is comment-level in this setting) representation together with higher granularity token-level representations to jointly train the network. We set the dimension of lower granularity embedding representation equal to the number of classes in the task. We jointly train tasks where number of classes are the same, that is, CL & TL fallacy identification tasks are trained together and so are CL & TL fallacy type identification tasks. We use sigmoid activation as it is the best model for their fragment (token) level classification and is comparable for the sentence level classifier. This model has been shown to give good results for predicting propaganda techniques, which include fallacies.

Conversation context. Our dataset is rich in textual information related to the COI, which could improve prediction. We define context as the parent comment of COI (if it exists, the string "None" otherwise) or the submission title. This is provided to the classifier in the format: `[CLS] COI Tokens [SEP] Context tokens [SEP]`. The `Context tokens` get a 'non-fallacy' token-level label at the training time, but during the validation or test set evaluation, only the COI token labels are used. The `[CLS]` token is used for CL tasks. This results in four extensions of the previous models: **BERT-T, BERT-P, MGN-T, MGN-P**, where T stands for title and P for parent comment.

Setup. We use PyTorch (Paszke et al., 2019) and the pre-trained BERT model (Devlin et al., 2019; Wolf et al., 2020). We fine-tune BERT using batch size 8, maximum sequence length 256 for COI & 64 for context, and monitored the macro-averaged F1[2] score on the validation set, as identification of all classes is equally important. We use the AdamW optimizer, with a learning rate of $5e^{-5}$. We weigh the cross-entropy loss function according to the class distribution in training data. We split the dataset into training (70%), validation (20%) and test (10%) sets, hence the full dataset has 6823, 1950 & 977 and annotated dataset has 2351, 671 & 336 comments respectively. We repeat the experiments with 5 different random seeds for the network intialization and we average the results.

5.2 Discussion

In Table 4, we show the results of comment level fallacy and fallacy type identification. All the results are macro scores (precision, recall and F1). The MGN models obtain the best results, most often when context is added. The full dataset pro-

[2] All reported F1 scores are macro F1.

Model	Binary			Multi class		
	P	R	F1	P	R	F1
Full dataset						
Random	47.67	47.67	47.67	9.98	10.02	10.00
BERT	66.31	66.28	66.15	50.03	48.80	48.30
BERT-T	67.54	69.01	67.99	46.93	49.49	46.57
BERT-P	68.73	68.75	68.52	38.08	49.85	41.83
MGN	69.50	70.01	**69.69**	47.87	48.59	47.14
MGN-T	70.73	68.76	69.61	51.18	48.22	**49.06**
MGN-P	71.15	68.72	69.62	50.06	50.38	48.53
Annotated data						
Random	46.35	46.35	46.35	13.01	13.16	13.04
BERT	66.918	64.00	61.16	62.25	55.63	57.83
BERT-T	66.76	66.57	66.44	62.03	55.88	57.93
BERT-P	66.72	66.54	66.45	61.08	56.61	57.90
MGN	67.72	67.54	67.45	59.81	54.72	56.19
MGN-T	69.57	69.27	**69.20**	62.72	55.91	**58.41**
MGN-P	69.53	68.99	68.86	62.96	55.85	58.17

Table 4: Comment level (CL) prediction for COI.

Model	Binary			Multi class		
	P	R	F1	P	R	F1
Full dataset						
Random	49.74	49.74	49.74	10.94	10.93	10.94
BERT	78.01	73.59	75.52	44.83	50.08	46.64
BERT-T	76.24	73.71	74.87	43.67	52.14	46.76
BERT-P	77.16	74.15	75.51	43.94	52.15	**47.07**
MGN	77.36	74.61	**75.86**	41.20	48.31	43.74
MGN-T	76.71	74.09	75.26	40.56	50.70	44.37
MGN-P	76.75	74.55	75.57	41.26	51.69	45.12
Annotated data						
Random	50.38	50.38	50.38	11.04	11.02	11.03
BERT	69.09	66.16	63.33	52.20	55.16	52.80
BERT-T	68.25	68.23	68.15	51.26	56.78	53.21
BERT-P	68.43	67.95	68.09	52.04	55.90	**53.44**
MGN	69.30	68.52	68.83	50.59	53.96	51.65
MGN-T	70.95	70.06	**70.26**	51.79	55.45	53.02
MGN-P	70.08	69.73	69.88	50.01	56.08	52.28

Table 5: Token level (TL) prediction for COI.

Fallacy	Full data		Annotated data	
	CL	TL	CL	TL
Appeal to authority	44.47	85.65	54.37	75.11
Appeal to majority	45.11	26.69	66.41	36.11
Appeal to nature	69.16	51.22	72.16	57.55
Appeal to tradition	56.92	55.08	66.08	60.43
Appeal to worse problems	35.31	20.81	43.89	30.73
Black-or-white	42.03	31.69	51.29	38.05
Hasty generalization	18.76	21.60	44.24	43.41
Slippery slope	39.54	37.68	57.37	55.57

Table 6: F1 score per fallacy from best classifiers.

vides a wider mix of topics via noisy negative samples and pronounces the class imbalance, closer to a real sample of Reddit conversations. Despite this, the classifier is able to learn across all four tasks.

Table 5 presents the results for token level fallacy and fallacy type identification. BERT models obtain better results for the multi class setting, while MGN for the binary setting. This is comparable with the results reported in Da San Martino et al. (2019), where the authors observe a smaller improvement in classification for the token level prediction using MGN.

Adding more context in the form of title or parent of the COI generally led to improved performance. While the results are slightly better when adding the title, the differences are small. We speculate that parent and COI provided a complete argument, making fallacy detection a bit easier.

In Table 6, we show the F1 score per fallacy class. Appeal to authority, nature, and tradition perform well ($F1 > 40\%$) across all four tasks. Hasty generalization has a rather poor performance; this can be attributed to this fallacy's general difficulty, given that the workers also had low agreement on this fallacy (Table 2). We observe that generally the comment level prediction task is easier than the token level prediction, which is expected due to the granularity difference.

Topical confounds. While fallacies might appear more frequently in discussions on certain topics, a fallacy detection approach should identify the underlying argument structure, and not just the presence of a topic. For example, we do not want to label all discussions about nature as appeal to nature fallacies. To identify if the classifiers are sensitive to topical biases, we use the approach presented in (Kumar et al., 2019). We compute statistically overrepresented tokens in each propaganda technique in the training set using log-odds ratio with Dirichlet prior (Monroe et al., 2008). We present the top 10 tokens per fallacy in Table 7. We observe that for appeal to authority, nature and tradition, the tokens are topically cohesive, as they revolve around notions of authority, nature and tradition. For the other fallacies, while it is intuitive why some words may be overrepresented, there is no clear topical cohesiveness. To verify that our classifiers learn linguistic patterns and not topics, we replace the top 30 tokens strongly associated with each fallacy (computed from the training set) with a special token in the test set. We evaluate only the comment level prediction, as results on the token level might be hard to interpret given that we replace tokens. We show the results in Table 8. We observe a large decrease in F1 score (more than 10% on the full data) for 2 fallacies:

Fallacy	Overrepresented tokens
Appeal to authority	medical, experts, expert, field, university, listen, degree, dr., professional, academic
Appeal to majority	majority, billion, reality, cult, christianity, followers, believe, nations, news, believed
Appeal to nature	animals, nature, eat, natural, meat, humans, food, species, killing, animal
Appeal to tradition	meat, years, marriage, history, eating, culture, vegan, thousands, tradition, ancestors
Appeal to worse problems	worse, world, problems, country, people, living, compared, dying, poverty, priorities
Black-or-white	pick, review, tax, gun, god, instead, absurd, profits, industry, paycheck
Hasty generalization	http, grew, friends, muslim, went, business, seen, grade, jesus, drivers
Slippery slope	government, slippery, slope, ban, stop, speech, remove, guns, line, start

Table 7: Top 10 tokens statistically overrepresented in each fallacy in the training set.

Fallacy	Full data	Annotated data
Appeal to authority	42.47	54.03
Appeal to majority	41.59	64.02
Appeal to nature	21.62	33.28
Appeal to tradition	41.80	49.20
Appeal to worse problems	27.25	32.66
Black-or-white	39.10	48.72
Hasty generalization	13.40	41.36
Slippery slope	34.76	54.44

Table 8: F1 score on comment level (CL) per fallacy after removing top 30 overrepresented words.

appeal to nature and appeal to tradition. A big drop in the F1 score on the full data is more significant than on the annotated data, as the classifier would have seen more negative examples containing the confounds. Given the observed decrease in F1 score for these fallacies, an important future direction is to annotate more discussions containing the overrepresented words to find a better quality negative set, i.e., non-fallacious comments on the same topics. We note that for the other fallacies, the models appear to learn more complex language structures as they are less sensitive to the removal of the overrepresented words.

6 Conclusion and Future Work

In this work, we present a methodology for mining and labeling fallacious comments in online discussions. We find frequent fallacy mentions on Reddit and the subreddits in which they are the most prevalent. We create a large corpus of annotated comments and experiment with several neural methods for classification. We explore methods that consider the context of the discussion, and we show that they give better results.

There are several exciting directions for continuing this work. First, using our methodology, we can annotate more comments for the eight fallacies we studied in this paper, we can improve the negative example set or explore other types of fallacies. Second, we can study another aspect of the discussion, the speech acts. According to the pragma dialectic theory, an argument is composed of several speech acts. Investigating if certain speech acts are more prevalent in fallacious discussions might lead to improved detection of fallacies. Lastly, in the pragma dialectic theory of argumentation, fallacies are violations of rules of critical discussion, for example, the fallacies we annotated violate two rules, as described in Section 3. Given the significant number of fallacy types, we believe that a hierarchical approach to their detection could prove more efficient: identifying if a conversation violates one of the ten rules of critical conversation, and then for that particular rule identifying the type of fallacy.

Acknowledgements

We would like to thank the ACL reviewers for their helpful feedback. We would also like to thank Meghana M. Bhat and Dravyansh Sharma for their helpful comments on the initial draft. This work was performed using HPC resources from GENCI-IDRIS (Grant 2020-AD011011614).

References

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 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), pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Frans H. van Eemeren and Rob Grootendorst. 1995. Argumentation, communication, and fallacies: A pragma-dialectical perspective. *Philosophy and Rhetoric*, 28(4):426–430.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. 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. Association for Computational Linguistics.

Hans Hansen. 2020. Fallacies. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2020 edition. Metaphysics Research Lab, Stanford University.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. 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)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (Î³) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Judith A. Sanders, Richard L. Wiseman, and Robert H. Gass. 1994. Does Teaching Argumentation Facilitate Critical Thinking? *Communication Reports*, 7(1):27–35.

Frank Scalambrino. 2018. *Psychologist's Fallacy: 100 of the Most Important Fallacies in Western Philosophy*, pages 204–207.

John R Searle. 1979. Expression and meaning: Studies in the theory of speech acts.

Douglas Walton. 2005. Justification of argumentation schemes. *The Australasian Journal of Logic*, 3.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online. Association for Computational Linguistics.

A Frequent Fallacies on Reddit

In this appendix, we review the most frequent fallacies on Reddit. Our goal is to understand how easy would be to annotate such fallacies, by looking at their definition and examples of how well Reddit users understand those definitions.

Let A,B,C,D be examples of persons, events, or actions. An argument consists of a standpoint S (known also as a claim or conclusion) and the supporting evidence (known also as the premises) for the standpoint. Let the person making/supporting the standpoint be referred to as the protagonist and the person disputing the standpoint as the antagonist. When referring to either protagonist or antagonist, we use the term party.

An argumentation scheme is a formalization of the relation between the evidence presented and the standpoint to be defended. Types of schemes:

- symptomatic argumentation: what is stated in the argument premise is an expression or a sign of what is stated in the conclusion.
- argumentation based on similarities: analogy between what is stated in the argument premise and what is stated in the conclusion.
- instrumental argumentation: argument and the conclusion are linked by a very broad relation of causality.

Fallacies are classified based on the argumentation rules they break out of the ten introduced in (van Eemeren and Grootendorst, 1995). Each fallacy is presented by giving all its possible name variations that link back to the same definition, its definition, and an example.

Rule 1. Parties must not prevent each other from advancing standpoints or casting doubt on standpoints.

Genetic fallacy. *Definition.* The antagonist rejects a claim stating that the source of the claim should not be trusted. The unexpressed premise is that every claim coming from the same source is likely to be false. *Example.* Fox News always writes junk news, I am sure that Hunter Biden did not break the law.

Ad hominem fallacy. *Definition.* The antagonist rejects a standpoint based not on the strength of the argument, but on perceived flaws of the protagonist, who is defending it. *Example.* You are such a bad student, I don't believe you got an A at maths.

Association / (guilt by/ honor by) association / reductio ad Hitlerum fallacy. *Definition.* The antagonist is disproving the claim of the protagonist by stating that this claim was supported by a bad group, hence the protagonist is also a bad person. *Example.* You say public healthcare is a good thing, but the communists say the same thing.

Tu quoque / appeal to hypocrisy / whataboutism fallacy. *Definition.* In this argument, the protagonist makes a claim S. The antagonist states that the claim S is in contradiction with the previous actions/attitudes of the protagonist (showing hypocrisy), thus the claim must be false. *Example.* To the statement "Putin is a killer", Trump responded, "There are a lot of killers. You think our country's so innocent?" - interview with Fox News' Bill O'Reilly.

Poisoning the well fallacy. *Definition.* This argument is a preemptive ad hominem, where the protagonist is attacked before advancing a standpoint. *Example.* I am sure Anna will say she gave the money back, but you know she always lies.

Rule 2. A party that advances a standpoint is obliged to defend it if the other party asks him to do so.

Argumentum ad ignorantiam / onus probandi / burden of proof / argument from ignorance / appeal to ignorance fallacy. *Definition.* The protagonist claims that a standpoint must be true because there is no or not sufficient evidence against it. As pointed out in (van Eemeren and Grootendorst, 1995), there can be two situations: *i*) the protagonist is challenging the antagonist to prove that their standpoint is wrong (rule 2) or *ii*) the protagonist is stating that because the negation of their standpoint cannot be proven true, then their standpoint is true (rule 9). *Example.* I have heard that vaccines are bad, prove me that they are good for your health!

Rule 3. A party's attack on a standpoint must relate to the standpoint that has indeed been advanced by the other party.

Straw man fallacy. The antagonist is: *i*) distorting the standpoint advanced by the protagonist (rule 3) or *ii*) attributing a false standpoint (rule 5). *Example.* Protagonist: I believe that women should have the right to abortion in the first term. Antagonist: So you're okay with killing babies.

Nirvana / perfect solution fallacy. *Definition.* In this argument, the protagonist is advancing the

claim that an action A is desirable as it will achieve a positive result. The antagonist rebuts this claim by stating that A will not achieve the perfect outcome, even if the perfect outcome is not specified in the claim. The antagonist modifies the claim, by stating "Action A will achieve the perfect outcome". *Example.* Protagonist: Using less plastic is good for the planet. Antagonist: We need to stop using plastic altogether to make any progress.

Moving the goalposts / raising the bar fallacy. *Definition.* This fallacy is similar to the nirvana fallacy, however, the antagonist is not aiming for the perfect outcome, but for better outcome than the one initially described by the protagonist. *Example.* We should stop killing animals for food, they feel pain. What about plants, how do you know if they don't feel?

Rule 4. A party may defend his standpoint only by advancing argumentation relating to that standpoint.

Ignoratio elenchi / irrelevant conclusion / missing the point fallacy. *Definition.* In this argument, the protagonist uses premises that are irrelevant to the claim. *Example.* His policies are not good enough, but my cousin says he talks well.

Rule 5. A party may not falsely present something as a premise that has been left unexpressed by the other party or deny a premise that they himself have left implicit.

Straw man fallacy. Already defined for rule 3.

Rule 6. A party may not falsely present a premise as an accepted starting point nor deny a premise representing an accepted starting point.

Circulus in demonstrando / petitio principii / begging the question / circular reasoning fallacy. *Definition.* In this argument, the evidence assumes that the claim is true. *Example.* Everyone likes me because I am the most liked politician.

Plurium interrogationum / fallacy of many questions / fallacy of presuppositions / complex question / loaded question fallacy. *Definition.* The standpoint brought forward by the protagonist is implying that at least another standpoint should be true. *Example.* Annie is a better person than that horrible guy John.

False dilemma / false dichotomy / bifurcation / black-or-white fallacy. *Definition.* The protagonist pushes the standpoint S that only the event or action A should be considered. The first premise is that only two events A and B are possible, when there is at least a third event C possible. The second premise is that one of the events is bad, for example B, thus only event A should be considered. *Example.* You must wear a mask each time you go out, otherwise you will die of COVID-19.

Rule 7. A party may not regard a standpoint as conclusively defended if the defense does not take place by means of an appropriate argumentation scheme that is correctly applied.

Relative privation / appeal to worse problems / not as bad as fallacy. *Definition.* The protagonist states that there exists A that is worse than B, therefore B is justified. The applied argumentation scheme is argumentation based on similarity. A and B are both bad actions, events or people, but instead of stressing the similarity, the protagonist tries to stress how A is bad, thus making B look better. *Example.* You shouldn't complain if the food is stale as there are millions of people starving who would be grateful for any meal they get.

Gambler's fallacy. *Definition.* The protagonist defends a probabilistic claim such as "an event A is very likely to occur". The mistake in argumentation appears if the evidence is based on falsely supposing that event A and event B are dependent, so if event A occurs, the probability of B occurring changes. This argument violates rule 7 as is it based on a faulty application of instrumental argumentation. *Example.* My coin landed twice in a row on heads, hence it should land next on tails.

Slippery slope / thin edge of the wedge / camel's nose fallacy. *Definition.* This fallacy consists in claiming that a small event A has a big unwanted consequence C. There is at least one more event B in the chain of causality (A will cause B, B will cause C), hence the slippery slope name of the fallacy. This argument violates rule 7, as the instrumental argumentation does not hold given that there is no clear causality chain between A and C. *Example.* If you break your diet and have one cookie tonight, you will just want to eat 10 cookies tomorrow and 20 the day after, and before you know it, you will have gained back the 15 pounds you lost.

No true Scotsman fallacy. *Definition.* The protagonist tries to make a generalization, which is a valid instrumental argumentation scheme: when a predicate P is true for an arbitrary member of a group, then it is true for any member of the group.

However, it changes the definition of the predicate P, so therefore the argument violates rule 7, as the instrumental argumentation does not hold anymore. *Example.* "No Scotsman puts sugar on his porridge".

Post hoc ergo propter hoc / temporal sequence implies causation fallacy. *Definition.* The protagonist states that because event A occurred first and event B occurred second, A caused B. This argument violates rule 7, as it tries to present an instrumental argumentation, without providing evidence that shows how event A and B are linked. *Example.* My boyfriend left me after he saw you, it must have been something you said.

Argumentum ad verecundiam / appeal to authority / argument from authority fallacy. *Definition.* In this argument, because the claim is supported by the opinion of a person with authority, then the claim is true. Rule 7 is violated because the symptomatic argumentation is incorrectly used: while authorities can make true claims, we can not consider them true as such if they are not backed up by evidence. *Example.* Being vegan makes no sense because my father said so.

Argumentum ad populum / appeal to widespread belief / bandwagon argument / appeal to the majority / appeal to the people fallacy. *Definition.* A claim is presented as true because many people believe it to be true. Rule 7 is violated because the symptomatic argumentation is not used correctly: while people do believe many things that are true, belief is not sufficient. *Example.* Being vegan makes no sense because so many of us are meat eaters.

Appeal to nature / naturalistic fallacy. *Definition.* The protagonist states that an action A is justified or good. The premise is that action A is good because it is natural. The argument violates rule 7 as it uses the symptomatic argumentation in a wrong way: some actions that are natural are good, however we cannot conclude they are good because they are natural. *Example.* Being vegan makes no sense as our body is designed for eating meat.

Argumentum ad antiquitatem / appeal to tradition fallacy. *Definition.* The protagonist states that an action A is justified or good. The premise is that it has always been considered as such in the past, but no further justification is given. The unexpressed premise of the argument is that everything that is done since a long time is good or justified as it has withstood criticism. However, this premise is also an opinion and not a fact. *Example.* Being vegan makes no sense as our ancestors have been meat eaters.

Divine / argument from incredulity / appeal to common sense fallacy. *Definition.* The standpoint appears incredible and not common sense from the perspective of the antagonist, and such it can be dismissed as false. In addition, everything that appears as common sense should be true. The argument uses the symptomatic argumentation in a wrong way: some actions that are incredible are false, however we cannot conclude that all incredible actions are false. *Example.* As disinfectant is efficient against Covid-19, it should be effective also if we drink it.

Hasty generalization fallacy. *Definition.* In this argument, the claim is supported by insufficient evidence through inductive generalization. More precisely, we know that predicate P is true for a sample of a population and we suppose it is true for the entire population. However, in this case the sample is either too small or it is not representative of the population. *Example.* The first two weeks of September were sunny, it means the rest of the month will be the same.

Volvo / anecdotal / proof by selected instances / person who fallacy. *Definition.* This fallacy is very similar to the hasty generalization, as a claim is not supported by sufficient evidence, but only a small set of examples. The difference between the two fallacies is that the examples in anecdotal fallacy are usually personal examples. *Example.* Two years ago when I visited Paris in September it was so nice and sunny, I am sure this year it will be the same.

Cherry picking / suppressed evidence / incomplete evidence fallacy. *Definition.* In this argument, a claim is backed by incomplete evidence, that is only a subset of facts that support the claim, while a large body of facts is overlooked. *Example.* My son is very smart, look he got an A at English! But what about all his bad grades before that?

Accident fallacy. *Definition.* The premises brought forward are generalizations that do not apply to the specific instances mentioned in the claim. *Example.* People bleed when they are ill, it means that your period is a sign of an illness.

Fallacy of composition. *Definition.* The claim is that a property P is true of a finite set S, also called in literature a whole. The evidence is that the property P is true for an element E that is part

of S. The unexpressed premise is that all the elements of the set are similar, which might be false and needs evidence. While this is similar to hasty generalization, in the latter there is no notion of a whole. Rule 7 is violated, as the instrumental argumentation does not hold. *Example.* Because the leaves of a tree are green, the tree is also green.

Fallacy of division. *Definition.* The fallacies of composition and division are the converse of one another. The claim is that something is true of an element E (let this be a property P), which belongs to a set S, called a whole. The evidence is that P is true for the set S. The unexpressed premise is that all the elements of the set are similar, which might be false and needs evidence. *Example.* If this tree is 100 years old, then each branch is 100 years old.

Argumentum ad temperantiam / argument to moderation / false compromise / middle ground fallacy / fallacy of the mean. *Definition.* Let S_1 and S_2 be two standpoints that represent very different opinions on the same topic. The claim is that a third statement, S_3, which is the middle point between the two, is true. *Example.* S_1 : We are having financial issues, we should fire all new hires. S_2 : No, we shouldn't fire any of the new people. S_3 : We should fire half of them.

Continuum / sorites / line-drawing / bald man fallacy / fallacy of the beard / fallacy of the heap. *Definition.* Let S_1 and S_2 be two extreme standpoints. Because there isn't a clear point where we pass from S_1 to S_2, it is supposed that there is no difference between them. *Example.* Once you drink a sip of alcohol you will become irresponsible and put your life in danger.

Rule 8. In his argumentation a party may only use arguments that are logically valid or capable of being validated by making explicit one or more unexpressed premises.

Special pleading fallacy. *Definition.* The protagonist applies rules or principles to other people or situations, but says this does not apply to the current situation without providing a justification. This is an application of a double standard. *Example.* While it is true he is only a teenager, I am sure he wasn't raped, he wanted to have intercourse with that woman.

Rule 9. A failed defense of a standpoint must result in the party that put forward the standpoint retracting it and a conclusive defense in the other party retracting his doubt about the standpoint.

Argumentum ad ignorantiam / onus probandi / burden of proof / argument from ignorance / appeal to ignorance fallacy. Already defined for rule 2.

Rule 10. A party must not use formulations that are insufficiently clear or confusingly ambiguous and he must interpret the other party's formulations as carefully and accurately as possible.

Equivocation fallacy. *Definition.* In this argument a word or expression is used with multiple meanings, thus trying to capitalize on the confusion to approve or disprove a claim. *Example.* If Americans are free, why do they have prisons? - here freedom has two meanings: the right to speak and act as one wants and the state of being imprisoned.

B Ethical Considerations

Worker compensation. Before assigning the tasks to crowd workers, the authors did several rounds of annotations themselves to determine the average time it takes to finish one HIT (10 fallacies). On an average it took about 20 minutes to annotate 10 fallacies. So we paid workers $5 per HIT, averaging to $15/hour. We still provided them 1 hour, in order to not put them under undue stress. Also, we did not request any personal information or opinions from the workers.

Banned and deleted content. Subreddits are closely monitored by the moderators. Users have to comply with Reddit's content policy, a lists a set of rules enforced by the admins on every community. Any rule violation (like bullying, use of hate speech, attacking marginalized or vulnerable groups, etc.) leads to the removal of posts/comments and, in some cases, banning a subreddit if the moderators fail to comply. The removal of such comments and posts ensures that we do not have any banned or deleted content in our dataset either.

Privacy of authors. None of our proposed methods does any profiling of Reddit users who made comments that appeared in our dataset. No identification of post/comment or their authors appears in our final dataset or input to the models.

Data quality. We describe our data collection process extensively in section 4. All the data samples appearing are annotated by two workers and resolved by authors if there is a disagreement between the workers.