# Persuade Me Not! Towards Understanding Persuasive Yet Fallacious Arguments

Paul Reisert †,‡ Kentaro Inui ‡,†
† RIKEN Center for Advanced Intelligence Project ‡ Tohoku University
paul.reisert@riken.jp inui@tohoku.ac.jp

## 1 Introduction

In argumentative dialogue such as debates and court cases, audiences are generally influenced by the more persuasive party. However, the more persuasive an argument, the higher chance of it being fallacious (i.e., consisting of logical flaws). For humans, such fallacious arguments can often times be overlooked, as more than 300 fallacy types exist [1].

Consider the following argument on the topic of *alcohol* consisting of a *claim* and its supporting *evidence*:

(1) **Claim**: *All people who drink alcohol are depressed.*
**Evidence**: *My friend drank daily and was never happy.*

For audiences unfamiliar with the topic, it may appear as though the argument is persuasive. However, for others, it can become readily apparent that the argument is a fallacious *hasty generalization* with reasoning such as “not all people can be considered depressed given one person’s situation.” It is crucial for humans, and especially machines, to identify such fallacious arguments for any given topic.

In the field of educational research, the usefulness of identifying fallacies as constructive feedback has been emphasized [2, 8, 7, 9]. In the field of NLP, previous works have addressed fallacy identification [4, 5]. However, no prior work has addressed providing specific constructive feedback for fallacious arguments which is increasingly important for applications such as student essays and debates, etc.

Towards enhancing a machine’s ability to recognize fallacious arguments, we aim to create a corpus which will allow us to model fallacious arguments and provide feedback for improving the argument (see Figure 1 for our overall goal). Ideally, a corpus for modeling fallacious arguments with reasoning should be large, contain many fallacious arguments along with their respective reasoning, and should be spanned across multiple topics.

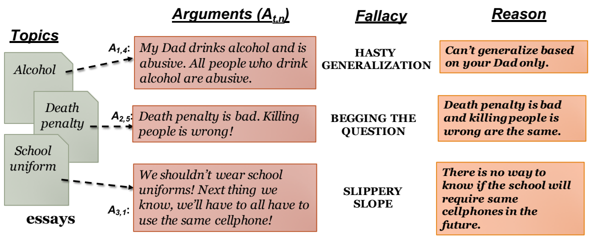


Figure 1: Overall goal of our work. We aim to automatically identify fallacious arguments spread across multiple topics and provide reasoning for improving the original argumentation.

In this work, we report our methodology for collecting fallacious arguments. We first leverage a popular online discussion forum for collecting posts with fallacious arguments and their reasoning. We then conduct an annotation study and a preliminary crowdsourcing experiment for identifying fallacious arguments. We discuss our results and future work towards creating a large-scale corpus of fallacious arguments and their fallacious reasoning.

## 2 Choosing a suitable collection of data

As aforementioned, a corpus for modeling fallacious arguments with reasoning should be large, contain many fallacious arguments along with their respective reasoning, and should be spanned across multiple topics. In this section, we describe a potential candidate domain for acquiring fallacious arguments and provide details.

### 2.1 Reddit

Similar to Habernal et al. [6], we utilize the online discussion forum Reddit[1] as a means for constructing our corpus. Reddit consists of, at the time of

[1] http://www.reddit.com


Figure 2: Interface for conducting our experiments. Annotators were first asked to identify the argument in the *DR* indicating a fallacy in the *OP*. Next, annotators were asked to select the region indicating the fallacious reasoning in the *DR*. Finally, annotators selected the fallacious argument in the *OP*.

| Fallacy | *DRs* |
|---|---|
| *begging the question* | 7,719 |
| *hasty generalization* | 1,850 |
| *slippery slope* | 114,869 |
| *straw man* | 39,789 |

Table 1: Total number of *DRs* after filtering via exact string match.

writing, millions of communities, referred to as subreddits (e.g., *news*, *Miyazaki*, *FanTheories*, etc.). On Reddit, users are able to make a unique thread consisting of a title and an original post (*OP*), and in response, other users can directly reply (*DR*) to the *OP* with a comment. Consequently, other users are able to reply to each DR, and so forth. For the purpose of this study, we focus on *OPs* and *DRs* in order to capture the original context of each thread. For collecting *DRs*, we utilize 14 years (12/2005 to 05/2019) of comments scraped and made publicly available.[2] In total, we acquire 5,743,794,806 comments. We filter out those immediately responding to an *OP* and collect 2,346,492,581 *DRs* (40.9%). For collecting *OPs*, we utilize PRAW [3], a Python Reddit API Wrapper.[4]

[2] `https://files.pushshift.io/reddit/comments/`
[3] `https://github.com/praw-dev/praw`
[4] Due to the large amount of *DRs*, we only collect *OPs* for *DRs* we filter in Section 3.

## 3 Annotation study

Given an *OP* and a *DR*, we would like to identify i) arguments in the *DR* which identify a fallacy in the *OP*, ii) the *DR*'s fallacious reasoning, and iii) the fallacious argument in the *OP*. Therefore, we first collect candidate *OP*/*DR* pairs and conduct a trial annotation and preliminary crowdsourcing experiment.

### 3.1 Collecting candidate fallacious *OPs*

For the purpose of collecting fallacious arguments and their reasoning, we filter *OP*/*DR* pairs by 4 common fallacy types [3]. We use an exact string match algorithm for filtering out pairs with one of the fallacies types in the *DR*. Shown in Table 1 are the fallacy types and*DR* containing the exact string match of the fallacy type.

For determining whether the pairs are easily annotable, we first tokenize each *OP* and *DR* and determine the average token length. We find that *OPs* and DRs, on average, have 173 tokens and 138 tokens, respectively. We also find the max number of tokens for *OPs* and *DRs* is 867,129 and 57,726, respectively.

### 3.2 Trial annotation

We conduct a trial annotation for determining the feasibility of collecting fallacious arguments and reasoning on top of Reddit. In preparation of large-

**Title**: "Regarding the NAP, we're going to have to admit there are flaws in it (re: spanking, abortion, meat-eating).

**OP**: [..] Raising a child is difficult, and even if one abstains from spanking (which i would support (abstaining from spanking, that is)), time outs (imprisonment), isolation (segregation), taking away toys (theft, if it was a gift), as well as denying icecream to a child (food regulation) and prohibiting them from watching 18+ movies (censorship); **we *have* to conclude that we use *force* against children in the process of raising them.** Excuses to the contrary are intellectually bankrupt. <br> <br> We *have* to be honest with ourselves, that morality *is not objective*. Morality is a tool, and the NAP is an excellent tool best applied to consenting adults. But this tool is *not* the most effective tool for all other situations. <br> <br> Discuss, but hopefully agree. And we don't have to worry about raging childish arguments, now that throwaway-o is gone (i used to love that guy. Sadface)."

**DR**: >we have to conclude that we use force against children in the process of raising them.
**This seems to me like a straw man. Those of us saying spanking is bad are not saying all uses of force against children are bad. Guardianship of a child includes some level of authority. Kids are not able to make certain decisions, and so their caretakers must make them for them. A kid might decide brushing their teeth is stupid, but their guardian should be able to force them to do it.** <br> <br> Maybe I misinterpreted you. <br> <br> >What objectively separates humans from animals? <br> <br> [..]

**Title**: Pro-lifers: if consent to sex means consent to *bearing* a child (no abortion), why doesn't consent to sex mean consent to *raising* a child (no adoption)?

**OP**: According to many pro-lifers, when women consent to sex, they thereby consent to (and commit themselves to) bearing any resulting children. And so, in deciding to having sex, these women have in effect *voluntarily waived their right to get an abortion*.\n\nNow, I find this pro-life claim utterly baffling: consent to sex is *clearly* different from consenting to anything further, many women deliberately use birth control to *avoid* pregnancy, many women plan on getting an abortion if they should end up pregnant, etc. According to this pro-life claim, it seems, we are supposed to interpret the act of consensual sex itself as involving some sort of mysterious *tacit consent* and *occult commitments* that are not only morally significant, but so overwhelmingly morally important as to *completely override the actual preferences of the woman*. I don't think actions carry occult commitments, and this all seems like superstition to me.\n\nBut here's my question. Let's suppose for the sake of argument that actions *do* carry occult commitments. Even granting this, we still need a way of telling what those commitments are. Without a method of interpretation, we're utterly in the dark. For example, a typical pro-lifer might say that the act of consensual sex carries the commitment to bear the child, waiving one's right to an abortion. **But a more radical pro-lifer might say that the act of consensual sex carries the commitment to bear *and raise* the child, waiving one's right to an abortion as well as one's right to put the child up for adoption.** My question is: how are we supposed to tell which interpretation is correct, and which occult commitments are (and are not) carried by the act of consensual sex?\n\n**EDIT**: After three hours, virtually every comment below is *completely missing the point*. Absolutely unbelievable, absolutely pathetic."

**DR**: **Equating abortion with adoption is really bizarre. It appears you've created a straw man.**

Figure 3: Examples of positive instances captured by our annotation study. *OP* fallacious arguments are shown in red, *DR* arguments indicating a fallacy are shown in green, and the fallacious reasoning is shown in blue. Note that we replace some text with [..] in order to reduce the example size.

| Criteria | Instances |
|---|---|
| *DR* indicates fallacy in OP? | 24/50 (48.0%) |
| *DR* contains fallacious reasoning? | 16/24 (66.7%) |
| *OP* contains fallacious argument? | 22/24 (91.7%) |

Table 2: Results from our annotation study for 50 *OP*/*DR* pairs on the topic of *abortion*.

scale annotation, we utilize ieturk[5], a Javascript-based, crowdsourcing-friendly annotation interface developed for the purpose of information extraction and named entity recognition. We modify the original interface to allow annotators to select boundaries for i) arguments in the *DR* which identify a fallacy in the OP, ii) the DR's fallacious reasoning, and iii) the fallacious argument in the *OP*. An example of the interface is shown in Figure 2.

We choose a random, controversial topic (i.e., *abortion*) and filter all *OP*/*DR* pairs with the controversial topic in the title. We randomly sample 50 pairs. The pairs were annotated by one annotator experienced in argumentation mining. The annotator first identified whether the *DR* indicated a fallacy in the *OP*. If not, they were asked to check the boxes shown in Figure 2. Otherwise, they highlighted the appropriate arguments in the text.

[5] `https://github.com/Varal7/ieturk`

The results of the annotation study are shown in Table 2. We observe that roughly half of the *DRs* indicate a fallacy in the *OP*. From these *DRs*, we observe that roughly 67% contain fallacious reasoning. Finally, if the *DR* indicates a fallacy in the *OP*, we observed that indeed most of the *OPs* contain a fallacious argument. Examples of pairs with a fallacious arguments and reasoning are shown in Figure 3.

### 3.3 Towards collecting fallacious arguments at a large-scale

To test the feasibility of collecting fallacious arguments at a large-scale, we conduct a preliminary crowdsourcing experiment. We use the crowdsourcing platform Amazon Mechanical Turk (AMT[6]). For worker qualification settings, we target workers who have completed 5,000 or more human intelligence tasks (HITs) and have an approval rating of 99% or more.

We originally set a reward of $0.20 per each completed HIT. Each pair was annotated by 3 crowdworkers using the ieturk interface. Similar to the trial annotation, workers were instructed to first identify whether the *DR* indicated a fallacy in the *OP*. If not, they were asked to check the boxes shown in Figure 2. Otherwise, they were asked to highlight the appropriate arguments in the text.

[6] `http://www.mturk.com`

| Segment | Exact | $\mathbf{P}_i$ | $\mathbf{P}_e$ |
|---|---|---|---|
| *DR* | 0.10 | 0.70 | 1.0 |
| *DR* reasoning | 0.0 | 0.29 | 1.0 |
| *OP* | 0.10 | 0.50 | 1.0 |

Table 3: Percentage of agreeing highlighted segments from crowdworkers in terms of exact, partial inclusive ($P_i$) and partial exclusive ($P_e$) string overlap.

Because we assume that crowdworkers are not familiar with all fallacy types, we train them on one fallacy type in our guidelines (for our experiment, we trained them on the *hasty generalization* type). In total, we annotated 10 pairs by all 3 annotators.[7] The average time for workers to complete one instance was 128 seconds, with a max of 433 seconds and a minimum of 14 seconds.

We report the Krippendorff's $\alpha$ of our results for the following: i) *DR indicates fallacy in OP*, ii) *DR contains fallacious reasoning*, and iii) *OP contains fallacious argument* as 0.44, 0.30, and 0.24, respectively. We then calculate pairwise string overlap between worker's highlighted segments using exact, partial (inclusive), and partial (exclusive) matching. The results are shown in Table 3. We observe that in all cases, there was always a partial overlap. In the case of *DRs*, roughly 70% contained inclusive overlap (e.g., "but hasty generalizations of things are deadly" and "I'm not trying to hate but hasty generalizations of things are deadly").

## 4 Discussion

Due to the nature of our approach, we are casting the task of fallacious argument identification as a single-label classification. However, in reality, an argument may consists of more than one fallacy type. Therefore, we must take this phenomena into account in our future work.

During our annotation study, we discovered that an exact string match resulted in noisy *DRs* (e.g., the term "slippery slope" can be used in a general context opposed to identifying a fallacy). Therefore, a more sophisticated filtering method must be applied to reduce the amount of negative samples. This is important when conducting a full-fledged crowdsourcing experiment, as annotating several negative samples can become costly.

During our crowdsourcing experiment, we originally employed 21 instances, but we discovered that only a fraction were annotated after 2 days. We attribute this to the fact that certain instances can have many tokens. In our future experiments, we will create a threshold of tokens based on our averages reported to ensure crowdworkers complete the task and are appropriately compensated.

[7] We originally employed 21 instances; however, we found most annotators stopped due to the large length of certain instances.

## 5 Conclusion

In this work, we proposed a method for collecting fallacious arguments and their reasoning. We first leveraged a popular online forum and collected candidate argument pairs. We then filtered the argument pairs by 4 common fallacy types and conducted an annotation study. From our results, we learned that fallacious arguments and their reasoning can be collected. To test the feasibility of annotating such pairs at a large scale, we conducted a preliminary crowdsourcing experiment and found that untrained annotators can reasonably identify fallacious arguments and their reasoning. In our future work, we will expand our argument pairs and conduct a full-fledged crowdsourcing experiment.

## References


[1] Bo Bennett. *Logically fallacious: the ultimate collection of over 300 logical Fallacies (academic edition)*. eBookIt. com, 2012.

[2] AE de Lima Alves. Constructive feedback. a strategy to enhance learning. *Medicina*, 68(1):88–92, 2008.

[3] Trudy Govier. *A practical study of argument*. Cengage Learning, 2013.

[4] Ivan Habernal, Patrick Pauli, and Iryna Gurevych. Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices. In *Proceedings of the Eleventh International Conference on LREC*, 2018.

[5] Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 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)*, 2018.

[6] 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 NAACL*, pages 386–396. Association for Computational Linguistics, 2018.

[7] Rohmani Nur Indah and Agung Wiranata Kusuma. Fallacies in english department students' claims: A rhetorical analysis of critical thinking. *Jurnal Pendidikan Humaniora*, 3(4):295–304, 2015.

[8] Witri Oktavia, Anas Yasin, et al. An analysis of students'argumentative elements and fallacies in students'discussion essays. *English Language Teaching*, 2(3), 2014.

[9] Yi Song and Ralph P Ferretti. Teaching critical questions about argumentation through the revising process: Effects of strategy instruction on college students' argumentative essays. *Reading and Writing*, 26(1):67–90, 2013.