Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4430–4438
Marseille, 20-25 June 2022

The Search for Agreement on Logical Fallacy Annotation of an Infodemic

Claire Bonial,[1] Austin Blodgett,[2] Taylor Hudson,[3] Stephanie M. Lukin,[1]
Jeffrey Micher,[1] Douglas Summers-Stay,[1] Peter Sutor[4] & Clare R. Voss[1]
[1]U.S. Army Research Lab, Adelphi, MD 20783
[2]Institute for Human & Machine Cognition, Pensacola, FL 32502
[3]Oak Ridge Associated Universities, Oak Ridge, TN 37831
[4]University of Maryland, College Park, MD 20742
claire.n.bonial.civ@army.mil

Abstract

We evaluate an annotation schema for labeling logical fallacy types, originally developed for a crowd-sourcing annotation paradigm, now using an annotation paradigm of two trained linguist annotators. We apply the schema to a variety of different genres of text relating to the COVID-19 pandemic. Our linguist (as opposed to crowd-sourced) annotation of logical fallacies allows us to evaluate whether the annotation schema category labels are sufficiently clear and non-overlapping for both manual and, later, system assignment. We report inter-annotator agreement results over two annotation phases as well as a preliminary assessment of the corpus for training and testing a machine learning algorithm (Pattern-Exploiting Training) for fallacy detection and recognition. The agreement results and system performance underscore the challenging nature of this annotation task and suggest that the annotation schema and paradigm must be iteratively evaluated and refined in order to arrive at a set of annotation labels that can be reproduced by human annotators and, in turn, provide reliable training data for automatic detection and recognition systems.

Keywords: logical fallacies, misinformation, inter-annotator agreement metrics

1. Introduction

With the outbreak of the COVID-19 pandemic, a parallel "infodemic" has emerged, defined by the World Health Organization as "too much information including false or misleading information in digital and physical environments during a disease outbreak."[1] We seek to identify such false or misleading information, or misinformation, in our efforts to develop a semantic search framework for COVID research, in which users can pose unconstrained natural language queries and receive a range of answers with explanations of their relevance (Bonial et al., 2020). Misinformation can take many forms and has become a quickly growing area of NLP research; here, we focus on an approach for annotating and automatically identifying *logical fallacies.* This area of research is particularly challenging because logical fallacies can be subtly encoded in the structure of a document across multiple sentences, making it difficult for human annotators or systems to recognize fallacies, yet also further motivating the need for assistance in finding such hidden, powerful sources of misinformation.

There are a growing number of relevant annotation schemas that could be leveraged for this task, many focused on the annotation of misinformation markers generally, including some logical fallacies (see §7). However, there are relatively few that are focused on logical fallacies as the primary phenomena of interest. After surveying existing schemas, we adopt the *Argotario* fallacy annotation schema of Habernal et al. (2017) and Habernal et al. (2018). This allows us to leverage the authors' existing dataset into a diverse training corpus for automatic identification of fallacies. Furthermore, we are able to explore differences in the realization of fallacies in the domain of COVID-19 documents, which is thought to be particularly prone to misinformation. Authors of the original corpus leverage a gamification approach to crowdsourcing the fallacy judgments (in their approach, one player first writes out a claim or argument, then indicates—as author of that claim—which fallacies are present, then a second player guesses at which fallacy the original author intended).

While crowdsourcing has a number of advantages in terms of cost and scalability, there is an open question as to whether the crowdsourced labeling of logical fallacies is in fact robust. We probe this issue by following a more traditional annotation approach, which allows us to evaluate our annotations as to whether or not the annotation schema category labels are sufficiently clear and distinct for reliable assignment across multiple annotators and documents. Two annotators and authors of this paper, each with formal linguistic training, annotate the sentences in a set of documents on six COVID-19 topics that are thought to be particularly rife with misinformation. Using our resulting corpus of 26 documents, we apply the Pattern-Exploiting Training (PET) procedure (Schick and Schütze, 2021) to train and evaluate automatic identification of the fallacies.

Our contributions here include the extension and application of the *Argotario* annotation schema within

[1] https://www.who.int/health-topics/infodemic

the domain of COVID-related texts (§3), the resulting novel dataset of COVID texts annotated with logical fallacies (§4), as well as an analysis of patterns in our corpus (§4.1), an evaluation measuring annotator agreement (§4.2), and a discussion of annotation challenges and sources of disagreement (§5). Additionally, we conduct a preliminary evaluation of the PET approach for automatically identifying fallacies (§6) using our annotated data. The results of our agreement analyses and evaluation of the automatic system demonstrate the challenging nature of this annotation task, suggesting that further refinement of the schema is needed before the fallacy types can be reliably distinguished by either humans or systems. This evaluation also illuminates important considerations for those adopting a schema or training data collected in a crowdsourced, "majority-rules" labeling approach generally: although the crowd "agrees" upon the same label for a particular annotation instance, this does not necessarily reflect a reliably replicable annotation decision that can be extended to new instances.

2. Background

The exploration of fallacies of focus in this paper fits into a broader research project on the development of an interactive information search system, distinct from typical question-answering systems in that users are able to present a full, unconstrained natural language question (as opposed to restricting their search to keywords) (Bonial et al., 2020). The goal is not to return a single answer in a one-off interaction, but rather to encourage an ongoing interaction between user and system to forage for the range of relevant answers, where these may differ with respect to focus, genre, as well as truth value and mis- or disinformation status. The ability of a system to detect and identify potential fallacies is extremely valuable in this envisioned interaction, as seen in Figure 1, where a system may answer a user's question—"Do I need to sanitize my mask?"—with both the answer as identified in a document, as well as a warning alerting the user to potential fallacies present, supplementing the sentences in the retrieved document. This exchange portrays our longer-term vision of how question-answering, information foraging, and mis- or disinformation detection can be unified under one framework.

3. Approach

Here we describe our approach to supporting fallacy detection in our framework, beginning with the fallacy schema adopted, annotation procedure, and corpus collected for annotation.

3.1. Fallacy Annotation Schema

So that we can leverage, add, and compare to their existing training corpus, we adopt the annotation schema of Habernal et al. (2017). We find their work to be uniquely valuable, as it is one of the few available schemas and datasets focusing on logical fallacies in

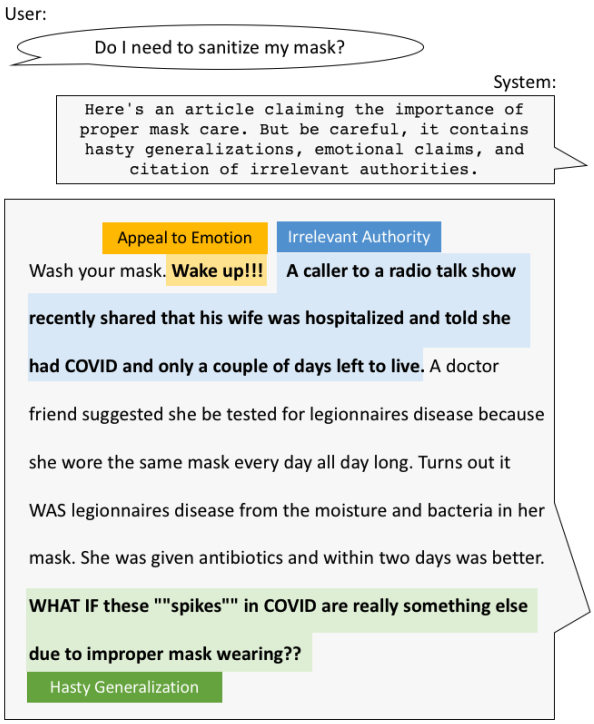

Figure 1: Envisioned exchange where the user's question is answered through dialogue, document retrieval, and fallacy annotation.

particular. The schema annotates just five fallacies: *Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization,* and *Irrelevant Authority*.

Thus, for each sentence of our document collection, annotators marked one of the following five choices, with labels and definitions adopted verbatim from the schema of Habernal et al. (2017); here, we have added examples and topics from our own corpus annotations:

1. Ad Hominem: The opponent attacks a person instead of arguing against the claims that the person has put forward. Example: *"It's just too convenient for vax-pimping scientists to claim that their precious vaccines don't work because not enough people are getting them"* (Topic: General vaccine safety and efficacy)
2. Appeal to Emotion: This fallacy tries to arouse non-rational sentiments within the intended audience in order to persuade. Example: *"It is time for families to wake up to uncloaking of the new world order in its glory."* (Topic: COVID-19 vaccine safety and efficacy)
3. Red Herring: This argument distracts attention away from the thesis which is supposed to be discussed. Example: *"Being a real scientist would be easy if it weren't for this 'needing evidence' stuff, just like being a professional golfer would be simple if it weren't for this 'having to put the ball in the hole thing."'* (Topic: SARS-CoV-2 virus origin)
4. Hasty Generalization: The argument uses a sample which is too small, or follows falsely from a

sub-part to a composite or the other way around. Example: (Preceding sentences: *"They're not reporting the number of deaths per million. In other words, they're not reporting the survivability rate."*) Annotation target: *"The answer here is don't mandate closures."* (Topic: Herd immunity)

5. Irrelevant Authority: While the use of authorities in argumentative discourse is not fallacious inherently, appealing to authority can be fallacious if the authority is irrelevant to the discussed subject. Example: *"'Inside Edition' also lauded Biden, Mitt Romney, and Tom Cruise for double masking recently."* (Topic: Mask safety and efficacy)

In our setup, as in the *Argotario* annotation procedure, a selection indicating "none" was a final annotation option. Note that this selection does not mean that no fallacy is present, but rather that none of the five fallacies of focus are present. To allow us to better understand overlap between the five fallacy labels versus potentially non-fallacious sentences in our agreement analyses, our final annotated corpus separates these labels into two levels. Each sentence annotation unit is accompanied by a "Level 1," two-way annotation of whether or not one of the five fallacies is present. If there is such a fallacy present, then there is a "Level 2," five-way annotation indicating which one is present.

3.2. Annotation Procedure

Here, we describe and contrast our own procedure with that of the original *Argotario* corpus.

3.2.1. *Argotario* Paradigm

The authors crowdsource the *Argotario* corpus of 1,344 snippets of English text with fallacy annotations. They use a gamification approach to data collection, in which the first player writes the snippet (ranging from a short sentence to a couple of sentences) including some claim and then indicates the "intended fallacy"—which of the five fallacies are present in their claim, or if none of those five fallacies are present. A second player is then presented with the first player's claim, and attempts to guess the first player's intended fallacy. The majority label given by second players is termed the "voted fallacy." The dataset includes information for each claim regarding the intended and voted fallacy, with an indication of how many second players voted. Instances that have five or more votes for a particular fallacy and that have entropy below a certain threshold according to the MACE metric (Hovy et al., 2013) are added to a gold standard subset.

3.2.2. Our Paradigm

In contrast to their crowdsourcing approach, we elected to extend fallacy annotation using the same five fallacies and their definitions, but relied on two linguistics-trained annotators to identify fallacies in documents related to COVID-19. This difference in procedure allows us to annotate fallacies in existing scientific journal papers, news reports, as well as social media posts about COVID-19. However, this change in procedure poses a new challenge: as we are now annotating fallacies "in the wild," we encounter a mixture of well-hidden fallacies that serve a particular author's agenda, as well as fallacies that may be entirely unintentional as a result of faulty reasoning. In both cases, the fallacy may only be clear in the broader context of the document, given the primary thesis or claim being made, the evidence presented (often in multiple sentences across the document) to support that claim, and how these claims and pieces of evidence relate to the social and cultural context. Detection may require implicit assumptions and knowledge held by readers of a similar socio-cultural background. This change in the annotation paradigm provides an opportunity for evaluating the annotation schema reliability using agreement metrics—an evaluation that was not reported for the original corpus collection and, as far as we are aware, was not used in the schema's development.

We also had to determine the unit of annotation for this task uniquely for our annotation paradigm. In the *Argotario* setup, the unit of annotation is the claim authored by the first player, who may choose to express the claim in a short sentence or several sentences. When translating this task to existing documents, some of the challenges mentioned above make determining the appropriate unit of annotation very difficult—a logical fallacy can be detectable within a single word (especially in the case of *Appeal to Emotion*), a clause or sentence, or, somewhat more commonly, as part of a set of sentences reflecting steps in reasoning where the fallacy is one step. Pilot experimentation demonstrated that leaving the unit open to interpretation resulted in vast disagreement in what should anchor the fallacy. For our purposes, we opted to separate each document into sentences and have both annotators annotate each and every sentence. The same sentence could be listed twice with different fallacies, where there were multiple fallacies exhibited in different parts of that sentence. Selecting this unit of annotation focuses the task and evaluation on the fallacy judgment, as opposed to the precise linguistic anchor of that fallacy, and also sets up our data nicely to serve as training prompts for PET.

Each individual document, with information on its genre and topic, was presented to two annotators (authors of this paper and native English speakers with linguistics training living in the U.S.). The document was presented as a spreadsheet, in which each sentence of the document was placed sequentially in its own row, and annotations were supplied in the adjacent column to each sentence instance within its row.

The annotation process took place across the period of about one month. There was an initial training period of about a week, during which annotators read and discussed the *Argotario* schema, annotated one of the COVID-related documents, and then discussed these annotations. Subsequently, after completing annotation of another 23 of the 26 documents independently ("first round" of annotation), the annotators met again

Topic	Genre (# of docs)
COVID Vax Safety	Online Medical Forum (2)
	Tabloid (1)
	Science Magazine (1)
	Social Media (1)
Herd Immunity	General News (3)
	Talk Radio (1)
Long Haulers	Online Medical Forum (4)
Mask Safety	General News (2)
	Social Media (3)
General Vax Safety, Efficacy	General News (1)
	Health Care Site (3)
SARS-CoV-2 Origin	General News (2)
	Scientific Article (2)

Table 1: Coverage of corpus topics and genres.

to discuss disagreements and established an agreed-upon gold standard subset of annotations for nine documents containing 226 annotation instances. This gold standard subset was used for training the PET model described in §6. After this, two more documents (44 annotation instances) were annotated, and this subset was used for a final inter-annotator agreement (IAA) measurement on the "second round" of annotations of our two annotators, and its gold standard annotations were used as a test set for the PET model.

3.3. Corpus Construction

The annotation documents in our corpus are related to the topic of COVID-19, largely from U.S. sources. Each article has one of six focus **topics** thought to be contentious and hence particularly rife with misinformation: mask safety, long haulers, herd immunity, general vaccination safety and efficacy, COVID vaccination safety and efficacy, and the origin of the SARS-CoV-2 virus. For each topic, we selected at least two documents reflecting opposing stances on the topic. For example, on the topic of herd immunity, two articles were chosen—one from Fox News and one from The New York Times. The Fox News article indicates that many states in the U.S. were already at the level needed for herd immunity to take place, so it was unnecessary for people to get vaccinated. The New York Times article describes how far off the U.S. was from reaching herd immunity, and without vaccines it would be impossible to reach.

We also paired the articles with different stances on a topic according to their **genre**. For example, two scientific articles are compared on the topic of the virus origin, where one argues for a man-made origin, while the other article argues for a natural origin. The resulting collection therefore allows for exploration of fallacies in documents demonstrating different perspectives on the same issues, and across different genres, while avoiding comparison between what we would expect to be very different genres with respect to fallacies, such as comparing social media posts to scientific articles.

Thus, the resulting corpus contains 26 documents that were manually selected based upon their main topic, the source genre, and the stance taken on the main topic.[2] The number of articles corresponding to a particular topic and genre are summarized in Table 1. Not all of our documents are full original texts; for example, the scientific journal sources only include the abstracts. The final corpus consists of 827 sentences. Of these, our gold corpus consists of 226 sentences, each of which has an adjudicated gold label.

Genre	%	Ratio
General News	49	41/83
Social Media	43	3/7
Health Care Sites	2	2/89
Medical Forum	6	2/31
Scientific Articles	31	5/16
Gold Corpus	23	53/226

Table 2: Percent of annotations with fallacies by genre in gold corpus. Ratio specifies the number of sentences.

4. Results

We provide an analysis and evaluation of the corpus.

4.1. Gold Corpus Analysis

Our objective in building and annotating the full corpus and the gold standard subset was to find actual examples of the five types of logical fallacies over which to develop and refine the annotation schema, not to make any generalized claims about the distribution of these types in the corpus. However, we hypothesized that the characteristics and hallmarks of the fallacies may differ across genres, and that there may be some patterns in the types of fallacies appearing in texts relating to COVID-19 that would be useful to both annotators and systems in recognizing these fallacies; thus, we offer observations on these patterns within the agreed upon gold corpus here.

Fallacies across genres The gold corpus included genres of: general news, social media, health care sites, online medical forums, and scientific articles (five of the full corpus' seven genres), where the majority of sentences came from general news and health care sites. The proportion of sentences with fallacies varied greatly by these genres, as shown in Table 2 for the gold standard corpus, made up of 226 annotations in total. The vast majority of sentences (77%), 173 of the 226 annotated sentences in the gold corpus do not contain any of the five fallacy types. This is not unexpected, given that our approach to corpus construction, unlike that of *Argotario*, did not involve actively eliciting fallacies, but rather to search and sample actual documents from the relevant topic space for the purpose of identifying logical fallacies via annotations, and this process yielded documents that did not contain fallacious claims.

The general news documents, including articles from Fox News, Reuters News, Time, and Forbes, had the highest fallacy percentage at 49%. Our small sample

[2]The corpus will be made available via data-sharing agreement pending publication.

Annotation	N	% Gold
Hasty Generalization	19	8.4
Appeal to Emotion	15	6.6
Red Herring	13	5.8
Ad Hominem	3	1.3
Irrelevant Authority	3	1.3
None of the Above	173	76.5
Total	226	100.0

Table 3: Percentage of annotations in gold corpus by fallacy type. N refers to the number of sentences.

of social media had a similar rate of 43%. Health care sites (e.g., Children's Hospital of Philadelphia website) had the lowest percentage of 2.2%, and the online medical forums (e.g., BMJ.com forum for health care professionals) was also quite low at 6%. The fallacy percentage of the scientific articles was 31%, which may seem surprisingly high on first glance, but one of the two scientific articles was a much-contested paper on the SARS-CoV-2 virus origin that was not peer-reviewed prior to publication on an open science site, where it is flagged as not following the norms of scientific rigour. Our single peer-reviewed scientific article contains 0 fallacies, while the non-peer-reviewed article has 5 fallacies for a fallacy percentage of 45%, so that taken together, they yield the high average fallacy percentage.[3]

Fallacy types The distribution of fallacies by type in the gold corpus is shown in Table 3. The most frequent fallacy type annotated was *Hasty Generalization*, which occurred 19 times, or 8.4% of the annotated sentences. The second most frequent fallacy annotated was *Appeal to Emotion*, which occurred 15 times and made up 6.6% of annotations. The next most frequent fallacy annotated was *Red Herring*, which occurred 13 times, making up 5.8% of the annotations. Both *Ad Hominem* and *Irrelevant Authority* were relatively infrequent in our gold standard corpus, both occurring three times each, and thereby each contributing to 1.3% of the corpus annotations.

4.2. Evaluation: Agreement Analysis

To evaluate the reliability of our annotations and schema, we compute IAA using Krippendorff's α (Krippendorff, 1980; Passonneau, 2004). Our choice of metric was motivated by the fact that the annotation categories are not equally distinct from one another, and form two levels of hierarchical tagsets (Artstein and Poesio, 2008; Artstein, 2017): first whether or not one of the fallacies is present, and if one is present, which of the five fallacies. To tease out the reliability of each level of annotation, we compute IAA across the Level 1 "two-way" judgment (i.e. *is one of the five fallacies present or not?*) and then the subsequent Level 2 "five-way" judgment (i.e. *which of the five fallacies?*). For Level 1 agreement, we simply compute IAA as to whether both annotators annotated that one of the five fallacies was or was not present. For Level 2 IAA, in only the subset of instances where both annotators agreed that a fallacy was present, we compute IAA as to whether annotators selected the same specific fallacy. IAA is summarized in Table 4.

We computed both levels of IAA for the full corpus, the first-round annotation portion of corpus, and the second-round/test portion of the corpus. This enables us to look for change over time after annotator discussion. We also computed Level 1 IAA for documents grouped by genre. Since the number of sentences when broken down by genre was small and so the number with fallacies even fewer, we did not compute Level 2 IAA within genre.

5. Discussion: Annotation Challenges

Overall, the IAA is quite low, demonstrating the challenging nature of this task even for annotators trained in linguistics who have exhibited reliable coding skills on several other complex annotation tasks. Although there is no absolute value for high agreement, values below .67 are thought to be inconclusive (Krippendorff, 1980).

The first-round corpus has the highest IAA of .54 for Level 2, while the second-round/test corpus has the highest IAA of .51 for Level 1. Thus, there is no evidence that the adjudication discussion establishing the gold standard corpus that occurred after the first round of annotation and before the second round of annotation of the test corpus led to any improvement in IAA, nor that there is any general improvement over time as annotators gain experience. Our annotators report that the second round/test set documents selected happened to be quite difficult.

Although our expectation was that Level 1 IAA (whether or not there is some fallacy present) would be higher than Level 2 IAA (which fallacy is present), this was not always the case. We posit that this may reflect a limitation of the choice to annotate at the sentence

Data	Level 1 IAA	Level 2 IAA
Full Corpus	.47	.51
First-round	.46	.54
Second-round/Test	.51	.31
General News	.23	-
Health Care Sites	.48	-
Online Med Forum	.26	-
Talk Radio	.48	-
Science Magazine	.31	-
Scientific Article	.33	-
Tabloid	.13	-

Table 4: IAA (Krippendorff's α) Annotator A1 vs. Annotator A2: across the full corpus, the first-round corpus, the second-round test corpus, and by genre.

[3] We note again that documents were selected on topics thought to be contentious and hence likely rife with misinformation; our samples are small, and we would not expect these genres to contain these levels of fallacies when treating other topics.

	HAST	EMOT	RED	AD	IRR
HAST	16	14	6	0	0
EMOT		40	13	3	1
RED			19	1	2
AD				3	0
IRR					2

Table 5: Confusion Matrix between annotators on Hasty Generalization (HAST), Appeal to Emotion (EMOT), Red Herring (RED), Ad Hominem (AD), and Irrelevant Authority (IRR), over 120 sentences where both agreed there was a fallacy.

level, thereby causing some disagreement in which sentence contained the fallacy in cases where the fallacious argument can only be identified as part of a broader context of multiple sentences setting up that argument, such as *Hasty Generalization*. We hypothesize that the annotation unit was a source of Level 1 disagreement since otherwise the two annotators found roughly the same number of fallacies in each document.

It does not appear to be the case that the genre, and whether or not the genre tends to include fallacies, has any influence on whether or not annotators can reliably identify these fallacies. General news articles have the highest fallacy rates in our gold standard corpus; while this genre does have one of the lower Level 1 IAAs, it is quite similar to the IAA of online medical forums, which had a very low fallacy rate.

Table 5 shows the confusion matrix between both annotators on the 120 sentences where it was agreed upon that a fallacy was present in the full corpus. The annotators agreed on the Level 2 fallacy label of 80 sentences, 67% of the corpus. Half of these agreements are *Appeal to Emotion* followed by *Red Herring* and *Hasty Generalization* with 24% and 20% agreement of Level 2 agreements respectively.

Taking a closer look, the most common disagreement cases are *Hasty Generalization x Appeal to Emotion* and *Red Herring x Appeal to Emotion*. Annotators noted that the definition of *Hasty Generalization*, as defined in the *Argotario* schema, seemed applicable to many claims that were not supported with adequate evidence or any evidence at all, making it a sort of catch-all category for subjectively "outlandish" claims. *Appeal to Emotion*, again as defined in the schema, was noted to be realized not only in rhetorical structure, but also in single, emotion-evoking words. As a result, it could co-occur with other fallacies. In such cases, annotators could list the same instance twice with two annotations; however, within the gold standard corpus the agreed-upon label was assigned, or in cases of disagreements, the agreed-upon label after discussion. The same sentence was never annotated twice with an agreed-upon label by both annotators, so there are no doubly annotated sentences in the gold standard.

Red Herring annotations could be particularly difficult, given that determining a *Red Herring* or distracting argument relies upon an awareness of the author's stance and main thesis arguing that stance. Our documents were carefully selected for their focus on certain topics, and these topics are provided to the annotator in a clear labeling of each document. Thus, the annotator need only determine the stance on that topic. Nonetheless, it is likely that cultural, implicit knowledge of the annotators plays a role in making assumptions about the stance/thesis of a document and, in contrast, any distracting, irrelevant claims.

Label HASTY GEN. vs. **Label EMOTION**
I suspect these draconian organized crackdowns on health freedom will become a permanent reality.
Government will use the Corona scare as a pretext to fast-track vaccine mandates into law everywhere.
Label RED HERRING vs. **Label EMOTION**
It is never wise to defer you personal decisions to an external authority, but especially now.
This dilemma has already redefined the landscape, giving rise to new authoritarianism.

Table 6: Disagreement examples relating to the topic of COVID vaccine safety.

Table 6 gives examples of these disagreements from documents relating to the topic of COVID vaccine safety. There is a highly emotional, dramatic nature to each of these examples that certainly makes *Appeal to Emotion* a plausible annotation, but there are also extreme conclusions being drawn that imply *Hasty Generalization*, as well as possible *Red Herring* arguments relating to authoritarianism as opposed to any direct evidence of vaccine safety.

Drawing from the common sources of annotation variation outlined in Artstein (2017), we identify two primary sources of variation that contribute to the disagreements that we observe. First, the fallacy labels have broad and overlapping definitions such that many circumstances arise where more than one fallacy label could apply. For example, *Hasty Generalization* and *Red Herring* fallacies can be written in such a way as to fit the definition of *Appeal to Emotion*, as seen in Table 6. Second, annotating these fallacies often requires applying external knowledge, especially subtle cultural knowledge, or making a subjective judgement, and these can vary from annotator to annotator. For example, our annotators reported differences in cultural knowledge that influenced decisions on whether or not a particular person cast as an authority should be considered an *Irrelevant Authority*.

6. PET for Automatic Fallacy Annotation

The *Argotario* corpus (specifically 430 instances manually checked by experts) has been used for classification experiments leveraging both a bi-directional LSTM model and an SVM model (Habernal et al., 2018). The results of those experiments demonstrated F1-scores ranging from 8 and 12% on the low end (corresponding to *Hasty Generalization* and Red Herring, respectively), up to 60% on the high end (for *Ad Hominem*). The authors attribute the relatively poor

Data / Agreement Level	A1 vs. A2	A1 vs. Gold	A2 vs. Gold	PET vs. Gold
Gold Corpus / Level 1	.46	.71	.82	-
Gold Corpus / Level 2	.54	.45	.56	-
Test / Level 1	.51	.77	.73	.22
Test / Level 2	.31	.72	.38	-.07

Table 7: Pairwise IAAs for Annotator 1 (A1) vs. Annotator 2 (A2), A1 vs. Gold, A2 vs. Gold, PET vs. Gold for both agreement levels on gold and test subsets.

performance on some fallacies to the small training corpus size, indicating that their approaches require large training corpora.

In our own evaluation as to whether the gold corpus can be used as training data for automatic identification, we leverage Pattern Exploiting Training described in (Schick and Schütze, 2021), which requires a relatively small amount of training data. PET is a method for tuning pre-trained language models to solve a linguistic task that the original models were not specifically designed to solve. PET is trained using a dataset composed of cloze-style (fill-in-the-blank) questions. As a template, we used the PET classifier, built on top of RoBERTa large, for the ReCoRD dataset (Zhang et al., 2018b), in which named entities in a passage are possible answers to a question. As multiple fallacies could occur within a single passage, ReCoRD conveniently allows us to either test for each target label one at a time or make a prediction from a set of target labels. In our experiments, the target labels are possible fallacies, and cloze-style questions asking which fallacy a passage demonstrates are attached to the passage, prompting the system to make a prediction: *Consider the passage [insert passage]. Which of the following fallacies is this an example of? a) Hasty Generalization, . . . The correct answer is [insert letter].*

We report our evaluation of the PET model for automatically annotating the five logical fallacies, or provide the appropriate "none of the above" annotation. In Table 7, we summarize IAA, again using Krippendorff's α, for the PET model trained on the gold corpus (consisting of 9 documents) and tested on the two held out documents in the test corpus. Here, we focus on IAA between PET and the gold standard label, and we also report for comparison the human IAA (repeated from Table 4), and the IAA between each of the human annotators and the gold standard label for the gold corpus and the test corpus.[4] The IAA for PET vs. the gold standard label on the test set is very low, dipping into negative values for Level 2 IAA, where negative values demonstrate systematic (beyond chance) disagreements. We note that our test set is small, so any tendencies here are just that, tendencies, and should not be taken as conclusive.

We acknowledge that because our training corpus identifies fallacies "in the wild" as opposed to eliciting fallacies of the five types through crowdsourcing as was done in the original *Agotario* paradigm, our data may have more subtle fallacies and will surely have less balance across the fallacy types. The latter was shown in Table 3, where we see that *Hasty Generalization* annotations outnumber others, but annotations finding no fallacy are by far the most numerous. Thus, we opted to run an additional evaluation where one author of this paper created a balanced training corpus by manually selecting what were thought to be five good examples from each category label from the *Argotario* corpus, resulting in a training corpus of 30 instances of fallacies. Using these 30 fallacies to train a second PET model, we tested this model on the same COVID-related test subset and again measure IAA between this model's output and the gold standard for the test subset. The IAA is notably better for the second model: .69 for Level 1 agreement and .09 for Level 2 agreement.

Although this result was surprising and further research is needed (perhaps comparing to a model trained on a carefully selected, balanced subset of our own data), we believe that this underscores the importance of considering the quality of the training corpus: when we can carefully curate data that reflects the clearly distinct, canonical examples of each category, then it is plausible our system can better replicate these distinctions. PET is, as its name implies, exploiting patterns in text, but fallacies, and especially those that are not elicited, may not have easily exploitable patterns. Instead, recognizing some kinds of logical fallacies may require the ability to follow the logical structure of an argument and see where it breaks down. Since fallacies may not be tied to easily exploitable patterns, it may be more important to have clear, canonical examples of fallacies than to have examples within a particular linguistic domain, such as the COVID domain.

7. Comparison to Related Work

There has been an explosion of activity in NLP on detecting misinformation and related tasks, including fake news detection and automatic fact-checking, stance and sentiment analysis, and rumor detection, resulting in various workshops and shared tasks (e.g, FEVER workshop). Thus, there are a variety of annotation schemas and datasets focused broadly on the de-

[4] We note that when comparing each annotator against the gold standard, we see comparatively high IAA of .71 and .82 for Level 1 annotation. This indicates that after the discussion period for establishing the gold standard corpus, the final, agreed-upon Level 1 label tended to agree with one of the annotator's original annotations. The somewhat lower Level 2 IAA when comparing to gold indicates that the discussion, at times, resulted in an agreed-upon gold label of which fallacy was present that was entirely new, or not either of the annotator's original labels.

tection and recognition of misinformation, which may have some overlapping categories with our research on fallacies, including the SemEval 2020 annotated dataset (Da San Martino et al., 2020a), and the credibility indicators outlined by Zhang et al. (2018a). Da San Martino et al. (2020b) offer a survey of relevant work on propaganda detection. We limit our comparison to the small, relevant slice of this expanding research area, which treats annotating and detecting logical fallacies in particular.

In addition to the *Argotario* corpus, Da San Martino et al. (2019) annotate a corpus for various propaganda techniques, including the annotation of 12 fallacies, of which only two categories overlap with the fallacies of interest in this paper: *Red Herring* and *Appeal to Fear and Prejudice*. As in our approach, the authors annotate journal articles as opposed to eliciting or seeking out particular fallacies. However, as a result, their corpus similarly suffers from an imbalance of fallacies that the authors conclude to be problematic for use of the corpus as training data.

In Sahai et al. (2021), potential fallacies are collected automatically from Reddit by searching for mentions of fallacies in comments, and then these are filtered through crowdsourced judgments. The only one of our five fallacies included in their schema is *Hasty Generalization*, for which the authors report the lowest IAA of any of their categories, measured via Cohen's κ, of .38. The highest IAA reported is .64 for *Appeal to Authority*. We note that our own overall IAA for Level 2 agreement across the full corpus is comparable to this range, .51, when using Cohen's κ. This underscores the challenge of this annotation task. The authors explore several models for automatic prediction of the fallacies, including BERT and MGN, with resulting F1 scores between 13 and 42% on the task most comparable to ours of labeling a comment with a particular fallacy. Unsurprisingly given the correspondingly low IAA, the lowest F1 score is for *Hasty Generalization*.

8. Conclusions & Future Work

Our application and evaluation of the *Argotario* logical fallacy schema has demonstrated the challenge of consistently recognizing these fallacies in documents relating to the COVID-19 pandemic. We conclude that, while a crowd-sourcing approach may establish that certain instances can be agreed upon by multiple annotators, this does not necessarily translate to evidence that schema distinctions can be reliably reproduced. If trained linguist annotators cannot reliably reproduce these distinctions, it seems unlikely that an automatic system trained on this data will be able to, and this is what our PET experimentation shows. Our error analysis illustrates problematic overlap in the three of five labels that are the most frequent in our corpus: *Hasty Generalization, Red Herring*, and *Appeal to Emotion*. Discussions surrounding these disagreements reveal the ways in which this fallacy annotation task brings together a nebulous combination of a small amount of linguistic knowledge along with larger amounts of social, cultural, and general knowledge. While annotators differ in these knowledge sources and in their weighting of how they contribute to an annotation, we question whether a computational approach to fallacy recognition would ever have access to these sources of information.

Yet the problems that could be addressed by an automatic system able to recognize fallacies continue to grow, making the need for solutions more urgent. Thus, the lessons learned in this process have led us to delineate a new misinformation annotation schema and guidelines, which we feel strongly must be developed and refined iteratively through rounds of piloting and IAA measurement in order to result in annotations that are of a reliable enough quality to serve as training data. Our evaluation here has demonstrated that much of the complexity of this annotation task arises from two main sources of variation: first, overlapping and unclear definitions of the fallacy labels, and second variation between annotators due to different levels of external knowledge or different subjective judgements. In our refined schema under development, we address the first source of variation by defining a taxonomy of fine-grained fallacy types with an associated annotation decision tree and criteria that define when each fallacy label applies. This approach has the advantage of providing a more detailed picture of the fallacies that appear in a document, as well as reducing the ambiguity of annotation decisions by use of a decision tree and criteria for fallacy labels that are non-overlapping. We address the second source of variation by excluding labels that consistently require external knowledge or subjective judgements (such as *Red Herring*).

Only when we have a reliable annotation schema and annotated corpus will we turn to considering the best computational approach for detecting and classifying fallacies. Although we will explore alternative approaches to classification, we will also continue to experiment with PET by first using a weighted approach. PET ranks the likelihood of each possible fallacy, and we noted that the correct answer was sometimes the second guess, with a relatively close difference.

The lessons we have learned in attempting to adopt and apply an existing annotation schema and training corpus are informative first for those interested in misinformation detection and recognition, for which it may be a particularly challenging task to develop a robust schema. However, we hope that what we have learned will also encourage any researcher making use of annotated training data to evaluate that data critically as part of determining what piece of the puzzle in their problem space may need to be improved: Are the annotation categories sufficiently clear and distinct so as to be reproducible, or is it an aspect of the computational approach that needs to be changed?

9. Bibliographical References

Artstein, R. and Poesio, M. (2008). Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Artstein, R. (2017). Inter-annotator Agreement. In *Handbook of Linguistic Annotation*, pages 297–313. Springer Netherlands, Dordrecht, Netherlands, June.

Bonial, C., Lukin, S., Doughty, D., Hill, S., and Voss, C. (2020). Infoforager: Leveraging semantic search with amr for covid-19 research. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 67–77.

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R., and Nakov, P. (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, November. Association for Computational Linguistics.

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R., and Nakov, P. (2020a). 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. International Committee for Computational Linguistics.

Da San Martino, G., Cresci, S., Barrón-Cedeño, A., Yu, S., Pietro, R. D., and Nakov, P. (2020b). A survey on computational propaganda detection. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization, 7. Survey track.

Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., and Gurevych, I. (2017). Argotario: Computational argumentation meets serious games. In *EMNLP (System Demonstrations)*.

Habernal, I., Pauli, P., and Gurevych, I. (2018). 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)*.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Krippendorff, K. (1980). *Content analysis: An introduction to its methodology*. Sage publications.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May. European Language Resources Association (ELRA).

Sahai, S., Balalau, O., and Horincar, R. (2021). 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. Association for Computational Linguistics.

Schick, T. and Schütze, H. (2021). Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Zhang, A. X., Ranganathan, A., Metz, S. E., Appling, S., Sehat, C. M., Gilmore, N., Adams, N. B., Vincent, E., Lee, J., Robbins, M., et al. (2018a). A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*, pages 603–612.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., and Van Durme, B. (2018b). Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.