ARL-TR-9641 ● FEB 2023

DEVCOM ARMY RESEARCH LABORATORY

Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE)

by Austin Blodgett, Claire Bonial, Taylor Hudson, and Clare Voss

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9641 ● FEB 2023

Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE)

Austin Blodgett
Institute for Human and Machine Cognition

Claire Bonial and Clare Voss
DEVCOM Army Research Laboratory

Taylor Hudson
Oak Ridge Associated Universities

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
February 2023	Technical Report	June 2021–September 2022

4. TITLE AND SUBTITLE
Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE)

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Austin Blodgett, Claire Bonial, Taylor Hudson, and Clare Voss

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-IC
2800 Powder Mill Road
Adelphi, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-9641

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This report details annotation guidelines for marking misinformation indicators, propaganda, and logical fallacies in text, culminating in the novel Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE) corpus. We give the motivation for developing this annotation schema, as well its relation to other parallel annotations. We describe the collection of a corpus of COVID-19 related texts to which the schema has been applied and supply sample annotations. We close with remarks on how this corpus is being leveraged to develop interactive question-answering and information extraction systems that are able to automatically detect and explain potential misinformation in queried documents.

15. SUBJECT TERMS
misinformation, propaganda, logical fallacy, annotation, natural language processing, Military Information Systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Claire Bonial
Unclassified	Unclassified	Unclassified	UU	88	19b. TELEPHONE NUMBER (Include area code) (301) 394-1431

Contents

List of Figures

List of Tables

1. Introduction and Research Background

The exploration of misinformation indicators in this report fits into a broader research project on the development of an 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). 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 misinformation indicators becomes paramount in this envisioned interaction, as seen in Fig. 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 (i.e., "Here's an article claiming the importance of proper mask care") as well as a warning alerting the user to potential indicators present, supplementing the sentences in the retrieved document. This exchange portrays our long-term vision of how question-answering, information foraging, and mis- or disinformation detection can be unified under one framework.

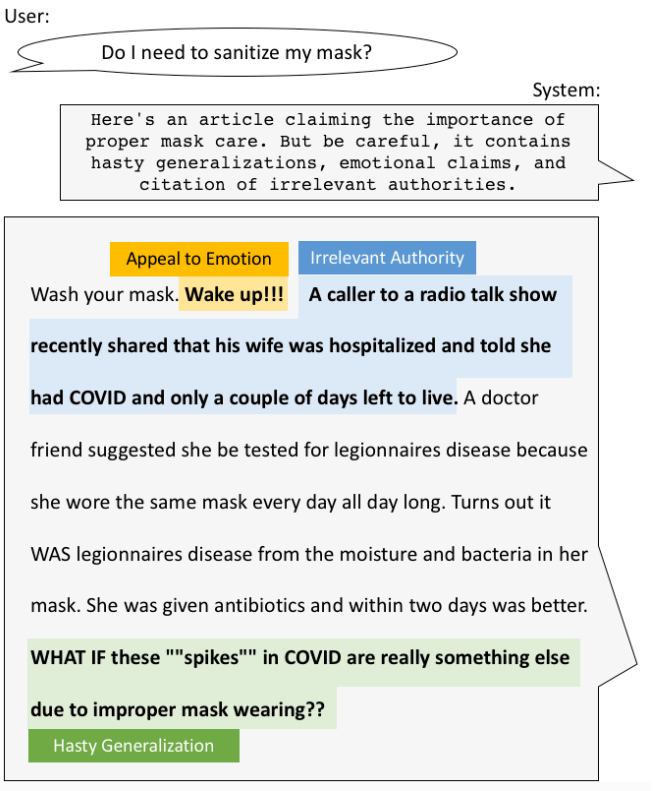

Fig. 1 Envisioned exchange where the user's question is answered through dialogue, the document is retrieved, and misinformation markers are highlighted and labeled

To support misinformation detection, we begin by annotating misinformation indicators in our domain of interest: scientific papers, general news, and talk radio, as well as health care sites and social media posts relating to COVID-19. We initially based our annotation schema on that of Habernal et al. (2017), which focuses on five logical fallacies: Ad Hominem, Appeal to Emotion, Red Herring, Hasty Generalization, and Irrelevant Authority. As a first step, we evaluated that existing annotation schema by measuring the ability of annotators to agree upon the categories across double-blind annotation of the corpus as well as evaluating the viability of the annotated data to serve as training data for automatic misinformation detection (Bonial et al. 2022a, 2022b). Our evaluation demonstrated that the schema categories were not sufficiently distinct to support high agreement rates among trained annotators, and this translated into poor automatic system performance. Thus, we began to develop our own schema, continuing to draw from existing annotation resources but evaluating iteratively for satisfactory agreement rates. Although we continue to refine the schema as we apply it to new data and measure agreement, the current resulting schema is described in the sections to follow.

The following sections of this report describe the Combined Annotations for Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE) schema and corpus. Section 2 describes the annotation schema with detailed instructions for annotators. Section 3 describes the CAMPFIRE corpus. Section 4 presents related work. Section 5 concludes and details future work. References are listed in Section 6. Appendixes A–C describe how to apply the annotation schema, the schema's history, and symbols and notation used throughout this technical report, respectively.

2. Annotation Schema

Here we describe the full annotation schema, beginning with our motivation for selecting certain kinds of annotation labels and distinct layers of annotation. After an overview of annotation practices relevant to all layers, we then detail annotation practices for each layer.

Sentence Annotations. In each annotation layer, each sentence in a document is considered a possible target of annotation. In the fallacy and rhetoric layers, sentences without an annotation are labeled "None." Layers are annotated independently of each other, so a single sentence might have one annotation in the fallacy layer and another annotation in the rhetoric layer. Using sentences as linguistic targets was a major design consideration, as it is not obvious as a matter of theory or practice whether fallacies or rhetoric are associated with a sentence, a

paragraph, a phrase, or a word. After annotation trials and repeated revisions and discussion, we found that using sentences as the annotation target reduced annotation ambiguity without significant loss of information.

Annotation labels and guidelines of the CAMPFIRE annotation schema are designed with three design principles in mind: *objective*, *explainable*, and *independent of external knowledge*.

Objective. Any features used for misinformation detection must be objective to be useful, both in the sense of being consistent across annotators and being impartial and removed from personal feelings and opinions. In terms of inter-annotator agreement, annotating fallacy consistently and objectively is a challenging task (Bonial et al. 2022a). To address this challenge, each label in the CAMPFIRE schema is given a clear description, and in the case of fallacies and rhetoric, each label is associated with examples and **litmus tests** for evaluating whether a given label should be annotated. CAMPFIRE litmus tests work by treating each fallacy or rhetorical technique as a semantic frame and annotators are asked to identify elements of that frame in a sentence before considering it for annotation. For example, the Slippery Slope fallacy contains elements corresponding to an *initiator,* a *starting event,* and a *resulting event* portrayed as the extreme result of the starting event. By checking for these frame elements, annotators can identify fallacies and techniques by relying on linguistic principles rather than intuition alone.

Explainable. Misinformation detection is most useful when users receive complete information and understand that information. As such, it is important that users not only understand that a fallacy or propaganda technique is present but why that fallacy or propaganda technique is misleading. To that end, CAMPFIRE organizes labels into fallacy and rhetoric taxonomies where each branch identifies why labels in that branch are misleading. For example, Inductive Fallacies include fallacies that use evidence improperly, jumping to a conclusion based on insufficient evidence or ignoring some evidence altogether. Each Inductive Fallacy has particular linguistic characteristics, but they are all misleading for similar reasons, so they are grouped together in our taxonomy.

Independent of External Knowledge. Fallacy and propaganda detection must be robust to new topics and information, which is fast changing in order to be useful. Because of that, detection needs to be independent of external knowledge, because external knowledge is not always available and can change as new information becomes available. To address this, CAMPFIRE was designed to only include annotation labels that could be identified without external knowledge. A number of annotation labels such as *Exaggeration* and *Red Herring* were removed from the

schema because annotation trials revealed that they could not be annotated consistently without some external knowledge.

2.1 Motivation

With the outbreak of the COVID-19 pandemic, a parallel "infodemic" has emerged, defined by the World Health Organization (Bradd 2021) as "too much information including false or misleading information in digital and physical environments during a disease outbreak." Thus, we seek to identify misleading or potentially problematic information (PPI) 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. PPI poses a problem not only for users seeking credible information, but also for natural language processing (NLP) tools such as document retrieval, question answering, and summarization, all of which involve retrieving text information from a large set of documents. When these NLP tools are designed with the assumption that all documents are credible and trustworthy or when there are no NLP tools available to distinguish PPI documents form trustworthy credible sources, these NLP tools will fail to inform users and may misinform them. Our research promoting the development of NLP tools for highlighting PPI for user scrutiny is based on detection of fallacies and propaganda techniques. Our long-term goal is to build information retrieval systems that not only retrieve answers based on a user's questions but also highlight fallacies and propaganda techniques in the retrieved text so that users can apply more scrutiny when a document might contain misinformation.

We identify two broad, equally important categories of NLP research on identifying text as misinformation: misinformation detection based on *internal features*, which includes propaganda detection and fallacy detection, and misinformation detection based on *external features*, which includes automatic fact checking. Automatic and partially automated fact checking are important to the promotion of credible information, but they require the development of external resources, such as relevant curated facts that can be checked against, before they can be used. In situations that are *fast changing*, such as the early days of a pandemic, a natural disaster, or other sudden event, misinformation tends to spread quickly and immediately before external resources can be developed for doing fact checking. Misinformation detection based on internal features only use information present in the text itself, such as whether the text relies on a known fallacy or propaganda technique. Thus, NLP tools for propaganda detection and fallacy detection has the potential to be applied sooner and be more robust to new data than NLP tools for automatic fact checking.

Our annotation schema includes three types of sentence-level annotation: annotation of motivation, annotation of fallacies, and annotation of rhetorical techniques. We base our categories for fallacies and rhetorical techniques on Habernal et al. (2017), who developed a schema of five fallacy labels, and Da San Martino et al. (2019), who developed a schema of 18 propaganda techniques. We further revised and expanded our own labels over the course of several annotation trials. We additionally developed decision trees and a strict set of rules called *litmus tests* for determining whether a fallacy or rhetorical technique was present in a sentence in order to promote annotator agreement. Additionally, we annotated what the most likely *motivation* the author had for each sentence based on a list of coarse labels.

2.2 Layer 1: Motivation

When trying to understand if a paper contains misinformation, one should search for the author's motivation behind each sentence to uncover their true intent. This task is meant to capture why the author uses a particular sentence or, in other words, what effect that sentence is intended to have on the reader. When reading an article that might contain misinformation, the annotators tasked with reading these documents look for these eight features in the motivation layer: *Persuasion*, *Attention*, *Distrust*, *Diversion*, *Concession*, *Advertisement*, *Entertainment*, and *Transition*. Each feature is labeled either "+" or "-".

Annotators might ask themselves during this task: Why is this sentence here? What is it meant to contribute to the document? Is this sentence used to _____:

1) Inform or Persuade (Persuasion +/-)
    a. Attempt to strengthen the author's thesis *(must provide information that is not redundant and that relates to the thesis).*
2) Grab Attention (Attention +/-)
    a. Attempt to increase the reader's level of attention.
3) Create Distrust (Distrust +/-)
    a. Attempt to lead the reader to distrust some target.
4) Divert or Confuse (Diversion +/-)
    a. Attempt to confuse the reader or direct their attention away from the author's thesis or information relevant to the author's thesis.
5) Make a Concession (Concession +/-)

    a. Acknowledge limitations of the author's claims or acknowledge counterarguments to the author's claims as valid.

6) Sell or Advertise (Advertisement +/-)

    a. Attempt to influence the reader to give money.

7) Entertain (Entertainment +/-)

    a. Attempt to entertain the reader.

8) Transition Sentences (Transition +/-)

    a. *This sentence only transitions between points and does not add any important information or emotion.*

If none of these features is positive, annotators should annotate a feature *Other* as "+" and add an explanation to the notes column.

2.3 Layer 2: Fallacy

In this phase of annotation, while reading a document, annotators are looking to identify logical fallacies that are present in the text. Fallacies are grouped into seven categories based on possible types of inference and which type of inference the fallacy is trying and failing to draw. Organizing the taxonomy this way also allows us to explain why techniques in this layer are fallacies, because we can compare them to credible forms and inference and easily identify the differences.

This layer is most like the work done in Argotario (Habernal et al. 2017), since logical fallacies are found more in text rife with misinformation. The following subsections describe the annotator instructions and decision tree and define the seven fallacy types.

2.3.1 Annotator Instructions

For each sentence in a document, annotators were asked to use one of the seven fallacy labels or none:

1) Seven different fallacy types—**Deductive Fallacy** (§1), **Inductive Fallacy** (§2), **Abductive Fallacy** (§3), **Testimony Fallacy** (§4), and **Rebuttal Fallacy** (§5).

2) **Decision Tree** (Section 2.3.2).

3) **Annotation Criteria**—Each fallacy label has a table of annotation criteria. They should read a few of these tables and understand them. They need to follow all rules and tests during annotation.

2.3.2 Decision Tree

Annotators use this decision tree to make annotation judgements for each sentence.

- Q1: Does this sentence evoke or rely on a fallacy? *(A fallacy can be anything that is used to draw a conclusion, present evidence, present testimony, present a hypothesis, or offer a rebuttal in a way that is logically, empirically, or methodologically flawed.)*
    - Choose Yes or No
        - Yes -> Continue to Q2.
        - No -> Annotate as **None**.
- Q2: Does annotating this sentence require external knowledge? *(External knowledge is any knowledge that cannot be inferred from the text itself and is not common sense knowledge. It includes knowledge that a sentence is false, knowledge about a source's qualifications, and knowledge about what constitutes a good analogy.)*
    - Choose Yes or No
        - No -> Continue to Q3.
    - Yes -> Annotate as **None**.
- Q3: Does the fallacy attempt to (choose from 1, 2, 3, 4, or 5):
    - 1) Draw a logical inference or use a formulaic argument? *(A fallacy like this should evoke a logical connective such as "and," "or," "if . . . then," "therefore," and one or more claims. See §1.)*
        - -> This is a **Deductive Fallacy**.
        - -> Identify the correct subtype from §1.
    - 2) Reason based on examples or evidence or based on the lack thereof? *(A fallacy like this should evoke evidence/data or the lack thereof. See §2.)*
        - -> This is an **Inductive Fallacy**.
        - -> Identify the correct subtype from §2.

- 3) Pose a hypothesis? *(A fallacy like this should evoke a hypothesis that is intended to explain some observations. See §3.)*
    - -> This is an **Abductive Fallacy**.
    - -> Identify the correct subtype from §3.
- 4) Rely on testimony from some source (possibly the author)? *(A fallacy like this should evoke a claim and a source or shared knowledge. See §4.)*
    - -> This is a **Testimony Fallacy**.
    - -> Identify the correct subtype from §4.
- 5) Offer a rebuttal? *(A fallacy like this should evoke a source or claim that is being rebutted. See §5.)*
    - -> This is a **Rebuttal Fallacy**.
    - -> Identify the correct subtype from §5.

- **If you cannot find a label that fits**: If an annotator cannot find a fallacy that fits any fallacy label in the taxonomy, they simply use the fallacy type label (or **Other** if there is no fallacy type that fits) and then add a description of the fallacy in the Notes column. This process helps to identify new fallacies and revise the taxonomy.
    - Choose 1, 2, 3, 4, 5, or 6:
        - 1) Annotate as **Deductive Fallacy**.
        - 2) Annotate as **Inductive Fallacy**.
        - 3) Annotate as **Abductive Fallacy**.
        - 4) Annotate as **Testimony Fallacy**.
        - 5) Annotate as **Rebuttal Fallacy**.
        - 6) Annotate as **Other**.
    - Write a description of the fallacy in the Notes column.

2.3.3 Deductive Fallacy

Definition

Each layer of fallacy labels has a higher label that can be used when the annotator is unable to decide on an annotation. Deductive fallacies can be compared to a

credible deductive inference. Deductive inference is the process of drawing a conclusion from a premise in a way that is logically sound. Deductive inference (or *logical inference*) draws conclusions by relying only on definitions of terms and logical expressions such as "and," "or," "not," and "if . . . then." Deductive fallacies attempt to draw a conclusion in a way that appears logical at first glance but fails to apply logical rules properly.

Deductive Inference

$X \Rightarrow Y \stackrel{\text{def}}{=} X$ *logically entails* Y

Examples of Deductive Inference:

- *If Roger is a cat, then Roger is an animal*. (True by definition)
- *We can either get a cat or not get a cat.* (These are the only two logical possibilities.)

Identifying a Deductive Fallacy

A Deductive Fallacy:

1) Evokes a logical connective such as "and," "or," "if . . . then," "therefore," or uses a formulaic argument.
2) Evokes one or more claims.
3) Is logically unsound.

Examples

- *If we don't get a cat, then we have to get a dog*. (Black and White Fallacy)

2.3.3.1 Fallacy 1.1: Begging the Question

Definition

Using the claim one is trying to prove as a premise.

Examples

- *Cats are the best because they're cats.*
- *Cats are the best because they just are.*

Non-Examples

- ✗ **Fails test 1:** *Cats are the best because they are fluffy.*
- ✗ **Fails test 1:** *Cats are the best because they have big claws.*

Notes

Also called assuming the conclusion.

Table 1 provides the annotation criteria for the Begging the Question Fallacy.

Table 1 Begging the Question Fallacy annotation criteria

Annotation Criteria		
	Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.	
Frame Elements	**Claim *X*:** The premise **Claim *X′*:** The conclusion	
	Test 1: $\boldsymbol{X} = X'$ (***X*** is logically equivalent to X', see non-examples)	
Logical Form	Premise : (Target) X	Ex: *Cats are the best pets.*
	Conclusion : X'	Ex: *Cats are the best pets.*

2.3.3.2 Fallacy 1.2: Black and White Fallacy

Definition

Asserting or presuming two claims or events to be the only available options.

Examples

- *If we don't get a cat, then we have to get a dog.*

Notes

Also called false dilemma or false dichotomy.

Table 2 provides the annotation criteria for the Black and White Fallacy.

Table 2 Black and White Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Event *E*:** **Event *F*:**	An event (often presented as more desirable) Another event (often presented as less desirable)	
Logical Form	Premise :	*E* and *F* are possible outcomes or options. **May be Implicit**: The premise of this fallacy is usually implicit.	Ex: *Getting a cat and getting a dog are both options.*
	Conclusion : (Target)	The only possible outcomes/options are *E* and *F*.	Ex: *If we don't get a cat, then we have to get a dog.*

2.3.3.3 Fallacy 1.3: Association Fallacy

Definition

This category includes fallacies that use a formulaic argument relying on biases and intuition in drawing a conclusion. Association fallacies attempt to conclude that something is good, bad, true, or false based on association with some characteristic such as *new* or *natural.*

Identifying an Association Fallacy

An Association Fallacy evokes an intuition or bias stemming from some cultural value or psychological tendency.

Example

- *We have to buy that new cat food they have.* (Appeal to Novelty)

2.3.3.3.1 Fallacy 1.3.1: Appeal to Nature

Definition

Asserting that something is better because it is natural (or bad because it is unnatural).

Examples

- *You have to feed cats raw meat because it's more natural.*

Table 3 provides the annotation criteria for the Appeal to Nature Fallacy.

Table 3 Appeal to Nature Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Event or Thing *A*:**	Event, state, or thing that is natural or unnatural, as presented by the author	
Logical Form	Premise : (Target)	*A* is natural_1 (or unnatural_2).	Ex: *Raw meat is natural.*
	Conclusion :	*A* is good_1 (or bad_2). **May be Implicit**: The conclusion will often be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Raw meat is better for cats.*

2.3.3.3.2 Fallacy 1.3.2: Appeal to Novelty

Definition

Asserting that something is better because it is new.

Examples

- *We have to buy that new cat food they have.*

Table 4 provides the annotation criteria for the Appeal to Novelty Fallacy.

Table 4 Appeal to Novelty Fallacy annotation criteria

<table>
<tr><th colspan="4">Annotation Criteria</th></tr>
<tr><td colspan="4">Rule 1: Make sure you can identify each frame element and the premise and conclusion.
Rule 2: Apply all the Tests listed under each frame element.
Rule 3: If an element or logical part is allowed to be implicit, it will be marked with May be Implicit along with instructions.
Rule 4: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the “Target” in the logical form.</td></tr>
<tr><td>Frame Elements</td><td>Event, Thing, or Claim A:</td><td colspan="2">Event, state, thing, or claim that is new or old, as presented by the author</td></tr>
<tr><td rowspan="3">Logical Form</td><td>Premise : (Target)</td><td>A is a new$_1$ (or old$_2$) thing/idea/practice.</td><td>Ex: A certain cat food is new.</td></tr>
<tr><td>Conclusion :</td><td>A is good/right$_1$ (or bad/wrong$_2$).</td><td rowspan="2">Ex: That cat food is better.</td></tr>
<tr><td></td><td>May be Implicit: The conclusion will often be implicit, but:
1. It must support the author’s thesis.
2. It must not replace an explicit conclusion if one is provided.</td></tr>
</table>

2.3.3.3.3 Fallacy 1.3.3: Appeal to Tradition

Definition

Asserting that something is better because it is old or status quo.

Examples

- *Old-fashioned cat food is better.*

Table 5 provides the annotation criteria for the Appeal to Tradition Fallacy.

Table 5 Appeal to Tradition Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Event, Thing, or Claim *A*:**	Event, state, thing, or claim that is new or old, as presented by the author	
Logical Form	Premise : (Target)	*A* is an old_1 (or new_2) idea/practice or is the status quo_1.	Ex: *A certain cat food is old or traditional.*
	Conclusion :	*A* is good/$right_1$ (or bad/$wrong_2$). **May be Implicit**: The conclusion will often be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *That cat food is better.*

2.3.3.4 Fallacy 1.4: Thought-Terminating Cliché

Definition

Use of a phrase or cliché to end discussion or imply that discussion is not meaningful.

Examples

- *It just is the way it is.*
- *I'm entitled to my own opinion.*

Table 6 provides the annotation criteria for the Thought-Terminating Cliché Fallacy.

Table 6 Thought-Terminating Cliché Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Claim *X*:**	The vacuous claim **Test 1**: *X* does not address the substance of issues being discussed.	
Logical Form	Premise : (Target)	Vacuous claim	Ex: *It just is the way it is.*
	Conclusion :	Further discussion would be unnecessary, pointless, or bad. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *We don't need to discuss further.*

2.3.4 Inductive Fallacy

Definition

Inductive fallacies can be understood by comparing them to a credible inductive inference. Inductive inference is the process of drawing a conclusion based on many observations. Inductive fallacies attempt to draw a conclusion from observations but do so while ignoring some examples or generalizing from too little information.

Inductive Inference

$E \Rightarrow F \stackrel{def}{=}$ *There are many observations of* $E \wedge F$ *and none (or few) of* $E \wedge \neg F$

or

$E \Rightarrow F \stackrel{def}{=} Pr(E \wedge F)$ *is high and* $Pr(E \wedge \neg F)$ *is low*

Identifying an Inductive Fallacy

An Inductive Fallacy:

1) Evokes evidence/data or the lack thereof.
2) Relies on that evidence/data in a methodologically flawed way.

Example:

- *No one has proven that cats can't understand humans.* (Appeal to Ignorance)

2.3.4.1 Fallacy 2.1: Hasty Generalization

Definition

Using information about an instance of a group or category to draw an inference about the entire group or category.

Examples

- *My cat is black, so all cats are black.*

Table 7 provides the annotation criteria for the Hasty Generalization Fallacy.

Table 7 Hasty Generalization Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Claim *X* :** a property **Category *A* :** group being generalized over		
	May be Implicit: ***A*** might be implicit		
	Member *a* : a thing or person in A ($a \in A$) such that ***X*** is true of ***a***		
Logical Form	Premise : (Target)	***X*** is true of a.	Ex: *My cat is black.*
	Conclusion :	***X*** is true of all members of ***A***. **May be Implicit**: The conclusion may be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *All cats are black.*

2.3.4.2 Fallacy 2.2: Appeal to Ignorance

Definition

Assuming the truth of some claim because it has not been proven false or the falsity because it has not been proven true.

Examples

- *No one has proven that cats can't understand humans.*

Table 8 provides the annotation criteria for the Appeal to Ignorance Fallacy.

Table 8 Appeal to Ignorance Fallacy annotation criteria

Annotation Criteria			
	Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.		
Frame Elements	**Claim *X*:**	Thing being argued for or against	
Logical Form	Premise : (Target)	*X* has not been $proven_1$/$disproven_2$.	Ex: *That cats can understand humans has not been disproven.*
	Conclusion :	*X* is $false_1$/$true_2$. **May be Implicit**: The conclusion may be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Cats can understand humans.*

2.3.4.3 Fallacy 2.3: Appeal to Accident

Definition

Ignoring or dismissing information as exceptional because it contradicts a conclusion.

Examples

- *Some people say cats are mean, but those are just the bad cats.*

Table 9 provides the annotation criteria for the Appeal to Accident Fallacy.

Table 9 Appeal to Accident Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Claim *X*:** The conclusion, refuted by A' **Evidence *A*:** All evidence for/against X		
	May be Implicit: A is often unstated. **Evidence A':** A subset of evidence A		
Logical Form	Premise :	A' suggests that X is false.	Ex: *Some people say cats are mean.*
	Conclusion : (Target)	A' is exceptional and not worth considering.	Ex: *Those are just the bad cats.*

2.3.4.4 Fallacy 2.4: Causal Fallacy

Definition

We can understand causal fallacies by comparing them to credible causal inference. Causal inference is the process of inferring a causal relationship between two or more events. Causal Inference is similar to Inductive Inference but has a higher burden of proof because of the difficult nature of proving causal relationships. Causal fallacies attempt to infer that a causal relationship exists without meeting the necessary burden of proof.

Note that for two events E and F, there are many possible relationships they could have with regard to cause and effect: E causes F (causation), E and F are unrelated (coincidence), F is caused by $E, E', E'', \ldots$ together (contributing factor), E and F are caused by G (confound), and so on. To identify the correct causal relationship between events, researchers need to do significantly more work than is required for simple inductive inference.

Causal Inference

We can say that E causes F if, while controlling for all other factors, E results in F and, and $\neg E$ results in $\neg F$.

Identifying a Causal Fallacy

A Causal Fallacy:

1. Infers a causal relationship between two events.
2. Is done in a methodologically flawed way (i.e., does not control for other factors or relies just on co-occurrence of events).

Example:

- *Many of the cat owners I know have asthma.* (Correlation–Causation Fallacy)

2.3.4.4.1 Fallacy 2.4.1: Correlation–Causation Fallacy

Definition

Inferring that A causes B if A happens when B happens.

Examples

- *Many of the cat owners I know have asthma.*
- *My neighbor has a cat and he has asthma.*

Notes

This is also called "cum hoc ergo propter hoc," which is Latin translating to "with this, therefore because of this."

Table 10 provides the annotation criteria for the Correlation–Causation Fallacy.

Table 10 Correlation–Causation Fallacy annotation criteria

Annotation Criteria			
Rule 1: **Rule 2**: **Rule 3**: **Rule 4**:	Make sure you can identify each frame element and the premise and conclusion. Apply all the **Tests** listed under each frame element. If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.		
Frame Elements	**Event *E*:** **Event *E′*:**	Initial event Accompanying event	
Logical Form	Premise : (Target)	E' often/sometimes happens with E.	Ex: *Many cat owners have asthma.*
	Conclusion :	E' is caused by E. **May be Implicit**: The conclusion may be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Owning a cat causes asthma.* (In reality, this could be because of coincidence, sampling bias, a confound variable, etc.)

2.3.4.4.2 Fallacy 2.4.2: Post Hoc Fallacy

Definition

Inferring that A causes B if B happens after A.

Examples

- *My cat has been upset ever since our neighbors moved next door.*

Notes

The full name of this fallacy is "post hoc ergo propter hoc," which is Latin translating to "after this, therefore because of this."

Table 11 provides the annotation criteria for the Post Hoc Fallacy.

Table 11 Post Hoc Fallacy annotation criteria

Annotation Criteria			
	Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.		
Frame Elements	**Event *E*:** **Event *E′*:**	Initial event Second event	
		Test 1: There is either an overt marker that E' occurs after E or that order of events is obvious from common-sense knowledge.	
Logical Form	Premise : (Target)	E' occurs after E.	Ex: *My cat became upset after our neighbors moved next door.*
	Conclusion :	E' is caused by E. **May be Implicit**: The conclusion may be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *My cat became upset because our neighbors moved next door.*

2.3.4.4.3 Fallacy 2.4.3: Slippery Slope

Definition

Asserting without evidence that allowing or causing some event will inevitably lead to a more extreme event.

Examples

- *If we allow pet cats, it's just a matter of time until someone has a pet alligator.*

Non-Examples

- ✗ **Fails test 1 and 2:** *If you smoke, you'll get cancer.* (Not every case of an event leading to a worse event is a Slippery Slope fallacy.)

Table 12 provides the annotation criteria for the Slippery Slope Fallacy.

Table 12 Appeal to Slippery Slope annotation criteria

Annotation Criteria			
	Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.		
Frame Elements	**Person/Group A:**	Initiator of the events $\boldsymbol{E}$ and $\boldsymbol{E'}$	
	Event E :	Starting event	
	Event E' :	Resulting event	
		Test 1: $\boldsymbol{E}$ and $\boldsymbol{E'}$ are intentional or presented as intentional.	
		Test 2: E' is presented as a more extreme version of $\boldsymbol{E}$**.**	
		Test 3: E is presented as an indirect cause of E', i.e. if E does not occur, E' is assumed not to occur.	
Logical Form	Premise :	$\boldsymbol{A}$ allows/causes event $\boldsymbol{E}$.	Ex: *Someone has a pet cat.*
	Conclusion : (Target)	$\boldsymbol{A}$ will allow/cause event $\boldsymbol{E'}$.	Ex: *Someone will have a pet alligator.*

2.3.5 Abductive Fallacy

Definition

Abductive fallacies can be understood by comparing them to a credible abductive inference. Abductive inference is the process of creating a hypothesis in order to explain some observations. While hypotheses are not guaranteed to be right, we can limit ourselves to only plausible hypotheses by following simple rules such as Occam's Razor. Occam's Razor says that our hypothesis should make as few assumptions as possible, while still explaining our observations. Hypotheses that violate this rule are abductive fallacies.

Abductive Inference

$E \Rightarrow Y \stackrel{\text{def}}{=}$ *Hypothesis Y is a possible explanation for observations E, Y accounts for all the observations and obeys Occam's Razor.*

Occam's Razor

A hypothesis should make as few assumptions as possible while explaining observations.

Identifying an Abductive Fallacy

An Abductive Fallacy:

1) Evokes a hypothesis that is intended to explain some observations (which may or may not be implicit).
2) Violates Occam's Razor or fails to explain the observations.

Example

- *The shortage of cat food is all because of immigrants*. (Scapegoat Fallacy)

2.3.5.1 Fallacy 3.1: Conspiracy Theory

Definition

Hypothesizing, without meeting a large burden of proof, that there exists a conspiracy orchestrated by some powerful group.

Examples

- *There is an evil, secret organization of people who want to kidnap our pet cats.*

Notes

Possible confusion with Appeal to Conspiracy.

Table 13 provides the annotation criteria for the Conspiracy Theory Fallacy.

Table 13 Conspiracy Theory Fallacy annotation criteria

	Annotation Criteria		
	Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.		
Frame Elements	**Person/Group *A*:**	Group claimed to orchestrate conspiracy **Test 1**: A is presented as an organized and powerful group with the aim of doing E.	
	Event *E*:	Conspiratorial act done by A, as presented by the author. **May be Implicit**: E may be unstated.	
	Event *F*:	Observations the existence of A and E is meant to explain.	
Logical Form		**May be Implicit**: F may be unstated.	
	Premise :	Observation: Some observations F. **May be Implicit**: The premise can be implicit.	Ex: *Cats sometimes go missing.*
	Conclusion : (Target)	Hypothesis: There exists a conspiracy orchestrated by powerful group A who is doing E.	Ex: *There is an evil, secret organization of people who want to kidnap our pet cats.*

2.3.5.2 Fallacy 3.2: Scapegoat

Definition

Hypothesizing, without meeting the burden of proof, that a particular person or group is entirely to blame for some complex event or harm.

Examples

- *The shortage of cat food is all because of immigrants.*

Table 14 provides the annotation criteria for the Scapegoat Fallacy.

Table 14 Scapegoat Fallacy annotation criteria

<table>
<tr><th colspan="4">Annotation Criteria</th></tr>
<tr><td colspan="4">Rule 1: Make sure you can identify each frame element and the premise and conclusion.
Rule 2: Apply all the Tests listed under each frame element.
Rule 3: If an element or logical part is allowed to be implicit, it will be marked with May be Implicit along with instructions.
Rule 4: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the “Target” in the logical form.</td></tr>
<tr><td rowspan="2">Frame Elements</td><td colspan="3">Person/Group A: Target of distrust
Event E: Event that is blamed on A</td></tr>
<tr><td colspan="3">Test 1: E must be complex enough that blaming it on a single person or a specific group requires burdensome assumptions (i.e., defies Occam’s Razor).</td></tr>
<tr><td rowspan="2">Logical Form</td><td rowspan="2">Premise :
Conclusion :
(Target)</td><td>Observation: E occurs.</td><td>Ex: There is a shortage of cat food.</td></tr>
<tr><td>Hypothesis: A is entirely to blame for E.</td><td>Ex: The shortage of cat food is all because of immigrants.</td></tr>
</table>

2.3.6 Testimony Fallacy

Definition

Testimony fallacies can be understood by comparing them to credible inferences from expert testimony or direct witness testimony. Credible inference from testimony requires that the source have credible and relevant expertise to speak about a topic or that the source be a direct witness of events, speaking only about what they can know as a direct witness. Testimony fallacies attempt to draw a conclusion based on testimony where the testimony might seem convincing but do not account for the credibility of the testimony.

Testimonial Inference

A reports $X \Rightarrow X$ if A has credible and relevant expertise related to X or is a direct witness of X.

Identifying a Testimony Fallacy

A Testimony Fallacy:

1) Evokes a claim.
2) Evokes a source or shared knowledge.
3) The source must either:
    a. Not have relevant expertise to or be a direct witness of the claim.
    b. Be unnamed or unidentifiable from the text.

Examples

- *It is known that cats can sense radio waves.*
- *We know that cats can sense radio waves.*
- *Studies show that cats can sense radio waves.* (no citations given)
- *Scientists say that cats can sense radio waves.* (no citations given)
- *Everyone knows that cats can sense radio waves.*
- *Ninety percent of people believe that cats can sense radio waves.*
- *My friend says that cats can sense radio waves.*

Non-Examples

- ✗ Cats can sense radio waves.

2.3.6.1 Fallacy 4.1: Bandwagon

Definition

Asserting as evidence that many people believe a claim in order to suggest that the claim must be true.

Examples

- *90% of people prefer cats.*

Table 15 provides the annotation criteria for the Bandwagon Fallacy.

Table 15 Bandwagon Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Group *A* :**	A population in which X is commonly believed, as presented by the author **May be Implicit**: The population *A* is allowed to be implicit.	
	Claim *X*:	Thing being claimed	
Logical Form	Premise : (Target)	X is commonly believed (among some group A).	Ex: *90% of people prefer cats.*
	Conclusion :	X is true. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Cats are better pets.*

2.3.6.2 Fallacy 4.2: Irrelevant Authority

Definition

Asserting as evidence that a specific person believes a claim where that person has no relevant knowledge or expertise for discerning the truth of the claim.

Examples

- *My spouse says that dogs are more intelligent than cats.*
- *I heard from a friend that cats can sense radio waves.*
- *Jane Doe, a well-known celebrity, says that cats can sense radio waves.*

Non-Examples

- ✗ **Fails test 1:** *I heard from a biologist that cats can sense radio waves.* (Sourceless Testimony)
- ✗ **Fails test 2:** *I heard from Jane Doe that cats can sense radio waves.* (no credentials given, so we cannot know whether this is a fallacy)
- ✗ **Fails test 2:** *I heard from a friend's neighbor that cats can sense radio waves.* (Sourceless Testimony)

Table 16 provides the annotation criteria for the Irrelevant Authority Fallacy.

Table 16 Irrelevant Authority Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:**	The source **Test 1**: This must be a single person or organization.	
		Test 2: *A* identified in the text by either (1) their name, (2) their credentials, or (3) their relationship with the author (friend, spouse, etc.). **Test 3**: *A* does not have specific expertise in *X*.	
	Claim *X*:	Common Examples: *a celebrity, a tv show, a friend or family member of the author* Thing being claimed by the source	
Logical Form	Premise : (Target)	Person A reports X.	Ex: *My spouse says that dogs are more intelligent than cats.*
	Conclusion :	X is true.	Ex: *Dogs are more intelligent than cats.*
		May be Implicit: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	

2.3.6.3 Fallacy 4.3: Sourceless Testimony

Definition

Relaying information that the author received without identifying a source. Note that if a relationship with the author is given (friend, relative, etc.) then the fallacy is Irrelevant Authority, not Sourceless Testimony.

Examples

- I heard *that cats can sense radio waves.*
- It is known *that cats can sense radio waves.*
- We know *that cats can sense radio waves.*
- A scientist said *that cats can sense radio waves.* (Without a citation this is not a named source.)
- Studies show *that cats can sense radio waves.* (no citations given)

- Scientists say *that cats can sense radio waves.* (no citations given)

Non-Examples

- ✗ **Fails test 1:** *I heard from a friend that cats can sense radio waves.* (Irrelevant Authority)
- ✗ **Fails test 1:** *The CDC says that cats can catch the flu.* (named organization)
- ✗ **Fails test 1:** *Scientists say that cats can sense radio waves. For example, According to Dr. Jane Doe . . .*

Table 17 provides the annotation criteria for the Sourceless Fallacy.

Table 17 Sourceless Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*** or **Event *E*:**	The unnamed original source of *X* or The shared knowledge of *X*	
		Test 1: ***A*** is not named in the text, and no relationship with the author (friend, spouse, etc.) or credentials are given.	
	Person/Group *B*:	Claimer who is not the original source of *X*	
	Claim *X*:	Thing being claimed	
Logical Form	Premise : (Target)	*B* reports having heard or read *X* (without naming the original source *A*).	Ex: *I heard that cats can sense radio waves.*
	Conclusion :	*X* is true.	Ex: *Cats can sense radio waves.*
		May be Implicit: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	

2.3.6.4 Fallacy 4.4: Appeal to Confidence/Disbelief

Definition

Asserting that the author is confident that a claim is true.

Examples

- *It's just obvious that cats are better.*

- *I am 100% sure that cats are better pets.*
- *Cats couldn't possibly be a good pet.*

Notes

We consider this fallacy to be a Testimony Fallacy where the testimony is from the author.

Table 18 provides the annotation criteria for the Appeal to Confidence/Disbelief Fallacy.

Table 18 Appeal to Confidence/Disbelief Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:**	The author(s)	
	Claim *X*:	Thing being claimed	
Logical Form	Premise : (Target)	The author *A* expresses their confidence that *X* is $true_1/false_2$.	Ex: *It's just obvious that cats are better.*
	Conclusion :	*X* is $true_1/false_2$. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Cats are better.*

2.3.6.4.1 Fallacy 4.4.1: Plain Folks Fallacy

Definition

Depicting a source as ordinary or plain in order to suggest that they are more trustworthy or correct. The plain person is generally depicted as being representative of the feelings and opinions of a large group. It is therefore a fallacy for the same reasons as the Bandwagon Fallacy.

Examples

- *You can trust me, I'm just an ordinary pet owner like you.*

- *I'm not a veterinarian, I'm just a regular person. So I know what it's really like to own a pet cat.*

Table 19 provides the annotation criteria for the Plain Folks Fallacy.

Table 19 Plain Folks Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person *A* :**	Source, usually the author(s), presented as a plain or ordinary **Common Phrases**: *A real person*, *a regular person, an ordinary guy, an average Joe*	
	Claim *X*:	Thing being claimed **May be Implicit**: ***X*** may not appear in this sentence or may be implicit.	
Logical Form	Premise : (Target)	The source *A* is presented as plain/ordinary	Ex: *I'm just an ordinary pet owner like you.*
	Conclusion :	X is true. or *A* is a trustworthy, representative source.	Ex: *My opinion is trustworthy and correct.*
		May be Implicit: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	

2.3.7 Rebuttal Fallacy

Description

Rebuttal fallacies can be understood by comparing them to a credible forms of rebuttal. The goal of credible rebuttal is to respond to the substance of an opponent's argument or scrutinize their sources or claims. There are a number of appropriate ways to do this, but rebuttal fallacies aim to discredit an opponent's claims in a way that might seem persuasive at first glance, but which is logically unsound or avoids the substance of the argument being disputed.

Credible Rebuttal

Rebutting an opponent's argument by

- presenting counterevidence,

- scrutinizing sources of information,
- providing an alternative explanation for observations,
- identifying inconsistencies in the opponent's argument,
- and so on.

Identifying a Rebuttal Fallacy

A Rebuttal Fallacy:

1) Evokes a source or claim, which is being rebutted.
2) Rebuts in a way that is methodologically flawed, such as critiquing something unrelated to the substance of an opponent's claims or credibility, misrepresenting an opponent's position, or relying on a conspiracy theory to discredit evidence.

Example:

- *People who like cats are just jerks!* (Rebuttal by Ad Hominem)

2.3.7.1 Fallacy 5.1: Appeal to Conspiracy

Definition

Dismissing counterevidence to the speaker's claims as lies fabricated by a conspiracy. Similarly, asserting that counterarguments or counterevidence to one's claim are "indoctrination" or "brainwashing" fit under this label.

Examples

- *People who like cats are brainwashed by the pro-cat shadow government.*

Notes

Easy to confuse with Conjuring a Conspiracy.

Table 20 provides the annotation criteria for the Appeal to Conspiracy Fallacy.

Table 20 Appeal to Conspiracy Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Group *B*:** **Claim *X*:**	Group claimed to orchestrate conspiracy Claim refuted by author as being fabricated by *B*	
Logical Form	Premise :	There is a conspiracy orchestrated by *B*.	Ex: *There is a conspiracy by a pro-cat group.*
	Conclusion : (Target)	Evidence for *X* is fabricated by *B*.	Ex: *The idea that cats are likable is fabricated by said group.*

2.3.7.2 Fallacy 5.2: Appeal to Cover-up

Definition

Stating that a source (or sources) does not mention X in order to suggest incompetence or malicious intent on the part of the source.

Examples

- *The news never tells you about all the people who were murdered by their cats.*
- *Doctors don't want you to know about the amazing health benefits of cat ownership.*

Discussion

While it is perfectly reasonable to critique a particular source or speculate about which evidence or information should be discussed, it is not generally reasonable to make assumptions about why some information was not mentioned, as there could be a variety of reasonable explanations: 1) It is possible that X is untrue, and that is why it was not mentioned, 2) It is possible that X is misleading or confusing without the presence of other information, or 3) It is possible that the source did talk about X, and the speaker simply missed that discussion. Related to Argument from Silence.

Table 21 provides the annotation criteria for the Appeal to Cover-up Fallacy.

Table 21 Appeal to Cover-up Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:** **Claim *X*:**	Target of distrust Claim not mentioned by *A*	
Logical Form	Premise : (Target)	Source A does not mention X.	Ex: *News sources don't mention anyone being murdered by their cat.*
	Conclusion :	A is incompetent or malicious for not reporting X. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Those news sources are untrustworthy, malicious, or incompetent.*

2.3.7.3 Fallacy 5.3: Rebuttal by Ad Hominem

Definition

Maligning a person rather than that person's claims, argument, or expertise in order to dismiss or discredit their position. This includes criticisms of a person's character, origin, or any trait of a person that is separate from their claims, argument, or expertise. This does not include insults in general unless an insult is used in place of a rebuttal.

Examples

- *I don't trust the opinion of a cat person.*
- *People who like cats are just jerks!*

Table 22 provides the annotation criteria for the Rebuttal by Ad Hominem Fallacy.

Table 22 Rebuttal by Ad Hominem Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:** **Claim *X*:** **Claim *Y*:**	Target of distrust Claim made by A, refuted by author Irrelevant criticism of A	
		Test 1: Y is not relevant to A's credibility, expertise, claim, or argument with respect to X.	
Logical Form	Premise : (Target)	Y is presented as true of A who reports X.	Ex: *Someone is a cat person.*
	Conclusion :	X is false or dubious. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Their opinion is untrustworthy.*

2.3.7.3.1 Fallacy 5.3.1: Rebuttal by Tone

Definition

Rebutting a person's claim by critiquing their tone or phrasing rather than the substance of their claims.

Examples

- *My neighbor is always making rude comments about not liking cats, and I don't take people who do that seriously.*

Table 23 provides the annotation criteria for the Rebuttal by Tone Fallacy.

Table 23 Rebuttal by Tone Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:**	Target of distrust	
	Claim *X*:	Claim made by *A*, refuted by author	
	Event *E*:	Tone or phrasing with which *A* says *X*	
Logical Form	Premise : (Target)	The tone/phrasing ***E*** used by *A* is disliked.	Ex: *My neighbor's comments about not liking cats are rude.*
	Conclusion :	*X* is false or dubious. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *My neighbor is wrong.*

2.3.7.3.2 Fallacy 5.3.2: Reductio Ad Hitlerum

Definition

Attempting to discredit a person's claim by making a superficial comparison to an evil or disliked person or group.

Examples

- *Cat owners are basically fascists.*

Notes

Also called guilt by association, a special case of false analogy.

Table 24 provides the annotation criteria for the Reductio Ad Hitlerum Fallacy.

Table 24 Reductio Ad Hitlerum annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:**	Target of distrust	
	Person/Group *B*:	Evil or disliked person/group	
		Common examples: Nazis, fascists, Adolf Hitler, Communists, Socialists	
	Claim *X*:	Claim made by *A*	
	Claim *X′*:	Claim made by *B*	
		May be Implicit: *X′* will often be implicit.	
		Test 1: *X′* is presented as similar to *X* by the author.	
Logical Form	Premise :	*B* believes *X′* **May be Implicit**: The premise will often be implicit.	Ex*: Fascist create strict rules to control their citizens, while cat owners also create rules for their pets.*
	Conclusion : (Target)	*A*, who believes *X*, agrees with or is the same as *B*.	Ex*: Cat owners are fascists (would agree with fascists in a substantive way).*

2.3.7.4 Fallacy 5.4: Straw Man Generalization

Definition

Misrepresenting a group's beliefs, often while posing a counterargument to the misrepresented belief, and priming listeners to perceive the author's position as more reasonable. For the sake of accuracy, we only annotated cases of this fallacy where the views of a heterogeneous group are misrepresented in an implausible way.

Examples

- *Dog lovers think that cats are evil!*

Non-Examples

- ✗ **Fails test 1:** *John Doe/The U.N. thinks that cats are evil.*
- ✗ **Fails test 2:** *Some dog lovers think that cats are evil.* (This is plausible because it doesn't generalize over the whole heterogeneous group.)
- ✗ **Fails test 3:** *Dog lovers think that cats are mammals.*

Table 25 provides the annotation criteria for the Straw Man Generalization Fallacy.

Table 25 Straw Man Generalization Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Group *A*:**	Group whose claim is misrepresented	
		Test 1: *A* is a (at least somewhat) heterogeneous group, not a person or organization.	
		Test 2: If *A* is quantified (e.g., *some, most, every*), the proportion of *A* must be a majority (i.e., *most* is allowed, *some* is not).	
	Claim *X*:	Claim believed by members of *A*	
	Claim *X′*:	**May be Implicit**: *X* will usually be implicit. A misrepresentation of *X*	
		Test 3: *X′* must not be derivable from common sense knowledge.	
		Test 4: *X′* is not equal to or entailed by *X*.	
Logical Form	Premise:	Group A reports X. **May be Implicit**: The premise will usually be implicit.	Ex: *Dog lovers think that dogs are friendlier than cats.*
	Conclusion: (Target)	Author reports/presumes that group A reports *X'*.	Ex: *Dog lovers think that cats are evil!*

2.3.7.5 Fallacy 5.5: Two Wrongs Make a Right

Definition

Assuming that some wrongful action is justified if the same or similar action was done by someone else.

Examples

- *People say I stole my neighbor's cat, but my other neighbor stole my dog from me!*

Notes

Easy to confuse with whataboutism.

Table 26 provides the annotation criteria for the Two Wrongs Make a Right Fallacy.

Table 26 Two Wrongs Make A Right Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group A:**	Wrongdoer being discussed	
	Person/Group B:	Wrongdoer who is an analogue of A	
	Event E: **Event E':**	Wrong done by A Wrong done by B	
		Test 1: E and E' are presented as similar	
Logical Form	Premise : (Target)	Person or group B did E'.	Ex: *My neighbor stole my dog!*
	Conclusion :	It is acceptable for person or group A to do E. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *It is acceptable for me to steal my neighbor's cat.*

2.3.7.6 Fallacy 5.6: Whataboutism

Definition

Responding to criticism by throwing the same criticism at other parties, often to imply that the critic is hypocritical, ultimately distracting from the original criticism without addressing it.

Examples

- *People say cats can be mean, but what about dogs?!*
- *People say cats can be mean, but have you heard about alligators?*

Notes

Easy to confuse with Two Wrongs Make a Right.

Table 27 provides the annotation criteria for the Whataboutism Fallacy.

Table 27 Whataboutism Fallacy annotation criteria

Annotation Criteria			
Rule 1: Make sure you can identify each frame element and the premise and conclusion. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions. **Rule 4**: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the "Target" in the logical form.			
Frame Elements	**Person/Group *A*:**	Wrongdoer being discussed	
	Person/Group *B*:	Wrongdoer who is an analogue of *A*	
	Event *E*: **Event *E′*:**	Wrong done by *A* Wrong done by *B*	
		Test 1: *E* and *E′* are presented as similar	
Logical Form	Premise : (Target)	Person or group *B* did *E′*.	Ex: *But what about dogs?!*
	Conclusion :	Concerns about *A* doing *E* are frivolous/hypocritical and should not be discussed. **May be Implicit**: The conclusion will usually be implicit, but: 1. It must support the author's thesis. 2. It must not replace an explicit conclusion if one is provided.	Ex: *Comments that cats can be mean do not warrant discussion.*

2.4 Layer 3: Rhetoric

In this phase of annotation, annotators identify rhetorical techniques that are present in the text of a document. Rhetorical techniques are categorized into five types based on their psychological effect on the reader. Organizing them in this way explains how rhetorical techniques work and why they can promote the spread of misinformation. This section uses similar labels that were in SemEval's persuasion techniques (Da San Martino et al. 2020a).

The following subsections describe the annotator instructions and decision tree and define the five types of rhetorical techniques.

2.4.1 Annotator Instructions

For each sentence in a document, annotators were asked to annotate a rhetoric label. They were asked to familiarize themselves with the following guidelines:

- The different rhetorical technique types—**Negative Emotion** (§1), **Positive Emotion** (§2), **Saliency Bias** (§3), **Memory Bias** (§4), and **Manipulating Behavior** (§5).
- **Decision Tree**
- **Annotation Criteria**—Each rhetoric label has a table of annotation criteria. Annotators were asked to read these tables and make sure they understood them. They had to follow all rules during annotation.

2.4.2 Decision Tree

Annotators used this decision tree to make annotation judgements for each sentence.

- Q1: Does this sentence evoke or rely on a rhetorical technique? (*A rhetorical technique can be anything that can be used to elicit an emotional response, make a message more salient or memorable, or manipulate a reader's behavior*.)
    - Choose Yes or No
        - Yes -> Continue to Q2.
        - No -> Annotate as **None**.
- Q2: Does annotating this sentence require external knowledge? (*External knowledge is any knowledge that cannot be inferred from the text itself and is not common sense knowledge. It includes knowledge that a sentence is false, knowledge about a source's qualifications, and knowledge about what constitutes a good analogy*.)
    - Choose Yes or No
        - No -> Continue to Q3.
        - Yes -> Annotate as **None**.
- Q3: What is the psychological effect of this rhetorical technique? Is it to:
    - Choose 1, 2, 3, 4, or 5:
    - 1) Elicit a negative emotion in the reader?
        - -> This is a **Negative Emotion** technique.
        - -> Identify the correct subtype. (See §1)
    - 2) Elicit a positive emotion in the reader?

        - -> This is a **Positive Emotion** technique.
        - -> Identify the correct subtype. (See §2)
    - 3) Make the reader more likely to pay attention by making a message more salient or dramatic?
        - -> This is a **Saliency Bias** technique.
        - -> Identify the correct subtype. (See §3)
    - 4) Make a message more memorable?
        - -> This is a **Memory Bias** technique.
        - -> Identify the correct subtype. (See §4)
    - 5) Prime the reader to be more likely to perform some specific behavior?
        - -> This is a **Manipulating Behavior** technique.
        - -> Identify the correct subtype. (See §5)
- **If you cannot find a label that fits**: If you find a rhetorical technique that does not fit any label in the taxonomy, you can simply use the rhetoric type label (or **Other** if there is no rhetoric type that fits) and then add a description of the rhetorical technique in the Notes column. This process helps us identify new rhetorical techniques and revise our taxonomy.
    - Choose 1, 2, 3, 4, 5, or 6:
        - 1) Annotate as **Negative Emotion**.
        - 2) Annotate as **Positive Emotion**.
        - 3) Annotate as **Saliency Bias**.
        - 4) Annotate as **Memory Bias**.
        - 5) Annotate as **Manipulating Behavior**.
        - 6) Annotate as **Other**.
    - Write a description of the rhetorical technique in the Notes column.

2.4.3 Negative Emotion

Definition

This category includes rhetorical techniques that attempt to elicit strong negative emotions in the reader, often to influence a reader's opinion or create a strong negative association. Appeals to emotion of this kind are not always fallacious and can be used by either credible or non-credible sources, but these techniques can indicate misinformation, especially in the absence of credible arguments.

If you cannot find the right label under Negative Emotion, also take a look at Appeal to Loyalty (§2c) and Flag Waving (§2c.1).

2.4.3.1 Rhetoric 6.1: Appeal to Anger

Definition

Use (or elicitation) of anger to support a claim. This can include sentences with an anger-evoking phrasing, content, or a sentence that suggests to the reader that they should feel anger.

Examples

- *Cats are killing the bird population.*
- *Do cat owners hate birds?!*
- *We should all be furious!*

Non-Examples

- ✗ **Fails test 1:** *I feel so angry.* (This reports the author's sentiment and does not have the reader as the experiencer.)

Table 28 provides the annotation criteria for Appeal to Anger Rhetoric.

Table 28 Appeal to Anger Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/ Group *A*:**	The experiencer/reader **Test 1**: The experiencer *A* corresponds to a reader of the author's intended audience. Sentences where the author expresses their own sentiment do not qualify (see non-examples).
	Event *E*:	**May be Implicit**: The reader may be unmentioned. Stimulus
		Note: May be inferred from context.
	Event *F*:	Feeling of anger
		May be Implicit: *F* may be indicated by particular lexical items or may be implicit.
		Test 2: Any appeal to emotion label can be emotion evoking either in it's phrasing or its content, but one of the following must be true: • **Phrasing**: The sentence's phrasing evokes emotion. Test: *Could this sentence be phrased more plainly?* • **Content**: The content of the sentence is designed to evoke emotion, and the sentence does not strengthen the author's thesis because either: ◦ The sentence adds only redundant information. ◦ The sentence adds information that doesn't strengthen the author's thesis.
Rhetorical Form Example	*[Cats are killing the bird population]E*. Implicit: This is intended to [anger]*F* the [intended reader]*A*	

2.4.3.2 Rhetoric 6.2: Appeal to Fear

Definition

Use (or elicitation) of fear or prejudice to support a claim. This can include sentences with a fear-evoking phrasing, content, or a sentence that suggests to the reader that they should feel afraid.

Examples

- *Cats often eat their owners.*
- *Are you next?!*
- *It gets even more frightening.*

Non-Examples

- ✗ **Fails test 1:** *I am feeling terrified.* (This reports the author's sentiment and does not have the reader as the experiencer.)

Table 29 provides the annotation criteria for Appeal to Fear Rhetoric.

Table 29 Appeal to Fear Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/ Group *A*:**	The experiencer/reader
		Test 1: The experiencer *A* corresponds to a reader of the author's intended audience. Sentences where the author expresses their own sentiment do not qualify (see non-examples). **May be Implicit**: The reader may be unmentioned.
	Event *E*:	Stimulus **Note**: May be inferred from context.
	Event *F*:	Feeling of fear **May be Implicit**: *F* may be indicated by particular lexical items or may be implicit.

		Test 2: Any appeal to emotion label can be emotion evoking either in it's phrasing or its content, but one of the following must be true:
		• **Phrasing**: The sentence's phrasing evokes emotion. Test: *Could this sentence be phrased more plainly?*
		• **Content**: The content of the sentence is designed to evoke emotion, and the sentence does not strengthen the author's thesis because either:
		  ○ The sentence adds only redundant information.
		  ○ The sentence adds information that doesn't strengthen the author's thesis.

Rhetorical Form Example	*[Cats often eat their owners]E*. Implicit: This is intended to [frighten]*F* the [intended reader]*A*	

2.4.3.2.1 Rhetoric 6.2.1: War Metaphor

Definition

Use of metaphor and imagery of war or battle, often with the intention of eliciting a passionate (and, in rare cases, violent) response from the listener.

Examples

- *Cat lovers should prepare to march into battle to defend your way of life.*

Table 30 provides the annotation criteria for War Metaphor Rhetoric.

Table 30 War Metaphor Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/Group *A*:**	The experiencer/reader
	Thing/Event *E*:	**May be Implicit**: The reader may be unmentioned. Actual event
	Thing/Event *E′*:	**May be Implicit**: E will often be implicit. Metaphorical description of E as warfare
	Test 1: E' must obviously not be literally true.	
Rhetorical Form Example	*[Cat lovers]A should prepare to [march into battle to defend your way of life]*E'. Implicit: This is metaphorical for some actual action E.	

2.4.3.3 Rhetoric 6.3: Appeal to Sadness/Pity

Definition

Use (or elicitation) of sadness or pity to support a claim. This can include sentences with an sadness-evoking phrasing, content, or a sentence that suggests to the reader that they should feel sadness or pity.

Examples

- *Hundreds of cats are euthanized every day. You could save one by getting a pet cat!*

Non-Examples

- ✗ **Fails test 1:** *I feel so sad.* (This reports the author's sentiment and does not have the reader as the experiencer.)

Table 31 provides the annotation criteria for Appeal to Sadness/Pity Rhetoric.

Table 31 Appeal to Sadness/Pity Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/ Group *A*:**	The experiencer/reader
		Test 1: The experiencer *A* corresponds to a reader of the author's intended audience. Sentences where the author expresses their own sentiment do not qualify (see non-examples).
	Event *E*:	**May be Implicit**: The reader may be unmentioned. Stimulus
		Note: May be inferred from context.
	Event *F*:	Feeling of sadness or pity
		May be Implicit: *F* may be indicated by particular lexical items or may be implicit.
		Test 2: Any appeal to emotion label can be emotion evoking either in it's phrasing or its content, but one of the following must be true: • **Phrasing**: The sentence's phrasing evokes emotion. Test: *Could this sentence be phrased more plainly?* • **Content**: The content of the sentence is designed to evoke emotion, and the sentence does not strengthen the author's thesis because either: ○ The sentence adds only redundant information. ○ The sentence adds information that doesn't strengthen the author's thesis.
Rhetorical Form Example	*[Hundreds of cats are euthanized every day]E*. Implicit: This is intended to [sadden]*F* the [intended reader]*A*.	

2.4.4 Positive Emotion

Definition

This category includes rhetorical techniques that attempt to elicit strong positive emotions in the reader, often to influence a reader's opinion or create a strong positive association. Appeals to emotion of this kind are not always fallacious and can be used by either credible or non-credible sources, but these techniques can indicate misinformation, especially in the absence of credible arguments.

If you cannot find the right label under Positive Emotion, also take a look at Saliency Bias (§3).

2.4.4.1 Rhetoric 7.1: Appeal to Optimism

Definition

Use (or elicitation) of positive emotions such as joy, hope, or optimism to support a claim. Also called wishful thinking, related to glittering generalities. This can

include sentences with a joy-evoking phrasing, content, or a sentence that suggests to the reader that they should feel joy or hope.

Examples

- *How could a cat hurt anyone? <u>They're too fuzzy and sweet</u>.*
- *If you get a cat as a pet, you'll sleep better, have a higher self esteem, and have a stronger immune system.*

Non-Examples

- ✗ **Fails test 1:** *I am so happy about this.* (This reports the author's sentiment and does not have the reader as the experiencer.)

Table 32 provides the annotation criteria for Appeal to Optimism Rhetoric.

Table 32 Appeal to Optimism Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/ Group *A*:**	The experiencer/reader
		Test 1: The experiencer *A* corresponds to a reader of the author's intended audience. Sentences where the author expresses their own sentiment do not qualify (see non-examples).
	Event *E*:	**May be Implicit**: The reader may be unmentioned. Stimulus
		Note: May be inferred from context.
	Event *F*:	Feeling of joy or hope
		May be Implicit: *F* may be indicated by particular lexical items or may be implicit.

		Test 2: Any appeal to emotion label can be emotion evoking either in its phrasing or its content, but one of the following must be true:
		• **Phrasing**: The sentence's phrasing evokes emotion. Test: *Could this sentence be phrased more plainly?*
		• **Content**: The content of the sentence is designed to evoke emotion, and the sentence does not strengthen the author's thesis because either:
		  ○ The sentence adds only redundant information.
		  ○ The sentence adds information that doesn't strengthen the author's thesis.

Rhetorical Form Example	*[They're too fuzzy and sweet]E*. Implicit: This is intended to make the [intended reader]*A* feel [happy]*F*.	

2.4.4.2 Rhetoric 7.2: Appeal to Flattery

Definition

Use (or elicitation) of flattery to support a claim.

Examples

- *You're so smart and caring that I bet you prefer cats over dogs.*

Table 33 provides the annotation criteria for Appeal to Flattery Rhetoric.

Table 33 Appeal to Flattery Rhetoric annotation criteria

Annotation Criteria	
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.	
Frame Elements	**Person/ Group *A*:** The experiencer/reader
	Test 1: The experiencer *A* corresponds to a reader of the author's intended audience.
	May be Implicit: The reader may be unmentioned.
	Claim *X*: Flattering statement
	Claim *Y*: Claim being tied to the flattering statement in this or an adjacent sentence.
Rhetorical Form Example	*[[You're]A so smart and caring]X that [I bet [you]A prefer cats over dogs]Y*.

2.4.4.3 Rhetoric 7.3: Appeal to Loyalty

Definition

Identifying support for a claim with loyalty to some group.

Examples

- *We cat lovers need to stick together!*

Table 34 provides the annotation criteria for Appeal to Loyalty Rhetoric.

Table 34 Appeal to Loyalty Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/ Group *A*:**	The experiencer/reader
		May be Implicit: The reader may be unmentioned.
	Group *B*: **Event *F*:**	Group Positive feeling of loyalty toward group *B*
		May be Implicit: *F* may be indicated by particular lexical items or may be implicit.
	Claim *X*:	Claim being associated with loyalty to the group
		Note: This rhetorical technique might involve a mix of positive and negative emotions because a positive feeling toward *B* might be used to encourage a negative feeling toward something else.
Rhetorical Form Example	*[We [cat lovers]B need to stick together]X!* Implicit: The [reader]*A* experiences [loyalty]*F* toward *B* .	

2.4.4.3.1 Rhetoric 7.3.1: Flag Waving

Definition

Implicitly or explicitly identifying support for a claim with patriotism or love of country.

Examples

- *People who love America love cats.*

Table 35 provides the annotation criteria for Flag Waving Rhetoric.

Table 35 Flag Waving Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Person/ Group *A*:**	The experiencer/reader
		May be Implicit: The reader may be unmentioned.
	Group *B*:	Country
	Event *F*:	Positive feeling of loyalty or patriotism toward country *B*.
		May be Implicit: *F* may be indicated by particular lexical items or may be implicit.
	Claim *X*:	Claim being associated with loyalty to or love of the country.
		Note: This rhetorical technique might involve a mix of positive and negative emotions because a positive feeling toward *B* might be used to encourage a negative feeling toward something else.
Rhetorical Form Example	*[People who [love]F [America]B love cats]X*. Implicit: The [reader]*A* who experiences loyalty toward *B*.	

2.4.5 Saliency Bias

Definition

This category includes rhetorical techniques that take advantage of natural human tendencies to place more focus on particular types of information. Saliency bias techniques attempt to make the reader pay more attention to a message by making it more salient (e.g., by making it seem more dramatic, vivid, or surprising).

If you cannot find the right label under Saliency Bias, also take a look at Appeal to Urgency (§5a) and Cliffhanger (§5b).

2.4.5.1 Rhetoric 8.1: Cliffhanger

Definition

Hinting at but withholding information in order to influence the reader to continue reading. May include rhetorical questions, prompting the reader to feel surprised, or directly prompting the reader to click or keep reading. This is a technique that in particular is associated with clickbait, but appears in other contexts as well.

Examples

- Rhetorical Question:
    - *Which pet is the nation's favorite? . . . It's cats.*
    - *Are there dangers to wearing a facemask?*

- Directly prompting the Reader to Click or keep reading:
    - *A certain someone murdered the mayor. Find out who.*
    - *Click here to find out . . .*
- Prompting the reader to feel surprise:
    - Y*ou won't believe . . .*
    - *It gets even stranger . . .*
- *I was shocked by what I'm about to tell you.*

Non-Examples

- ✗ **Fails test 1:** *It's shocking to learn that the mayor was murdered.*
- ✗ **Fails test 2:** *Someone murdered the mayor.*
- ✗ **Fails test 3:** *It gets even more frightening . . .*

Table 36 provides the annotation criteria for Cliffhanger Rhetoric.

Table 36 Cliffhanger Rhetoric annotation criteria

Annotation Criteria	
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.	
Frame Elements	**Person/ Group *A*:** The experiencer/reader
	May be Implicit: The reader may be unmentioned.
	Claim *X*: Information that is hinted at but withheld to keep the reader's attention. **Test 1**: The information being hinted at cannot be stated in the same sentence. (May or may not be answered in the article)
	Test 2: The sentence is written to keep the reader's attention, not to just provide information.
	Test 3: If there is a stronger emotion expressed, such as appeal to fear or appeal to anger, use that label instead.
Rhetorical Form Example	*[You]A won't believe [what this cat did while its owner was away]X.*

2.4.5.2 Rhetoric 8.2: Dramatization

Definition

Framing information in an unnecessarily provocative or dramatic way.

Examples

- *My neighbor's cats run around terrorizing the neighborhood.*

Table 37 provides the annotation criteria Dramatization Rhetoric.

Table 37 Dramatization Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Event *E*:**	Dramatized event
	Event E':	**May be Implicit**: E will generally be implicit. Dramatic metaphorical description of E **Test 1**: E' must obviously not be literally true.
Rhetorical Form Example	*My neighbor's cats run around [terrorizing]E' the neighborhood.* Implicit: An actual event E (such as *bothering*)	

2.4.5.3 Rhetoric 8.3: Vividness

Definition

Describing something in vivid detail in order to grab attention, especially of hypotheticals. This includes the use of specific anecdotes to draw the reader's attention.

Examples

- *Imagine a cat owner who comes home from a hard day of work at the oncology ward or the police station or the automobile factory and sees their fuzzy friend lying on the couch waiting all this time just to cheer their owner up.*

Table 38 provides the annotation criteria for Vividness Rhetoric.

Table 38 Vividness Rhetoric annotation criteria

Annotation Criteria		
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.		
Frame Elements	**Thing/Event *E*:**	Thing or event depicted in a vivid way
	Things/Events *E′*:	**May be Implicit**: *E* will generally be implicit. Vivid descriptions of *E* **Test 1**: *E′* must include information that is not necessary to the argument or message, but which is only included to add detail and saliency.
Rhetorical Form Example	*Imagine a cat owner [who comes home from a hard day of work . . .]E′.* Implicit: An event *E* (e.g., *A cat owner comes home and is happy to see their cat*).	

2.4.6 Memory Bias

Definition

This category includes rhetorical techniques that take advantage of natural human tendencies to remember particular types of information more readily. Memory bias techniques attempt to influence which information stays in a reader's memory (e.g., by using repetition or catchy slogans).

2.4.6.1 Rhetoric 9.1: Repetition

Definition

Repeating a previous sentence or part of a sentence for emphasis. For this label to apply, the sentence must not add any new information and must be obviously the same information (either word for word or slight paraphrase) expressed in a previous sentence.

Examples

- *Cats are the best pet . . . Cats are the best.*
- *Florida is 172 deaths per million—172?*
- *Texas amazingly, amazingly has the lowest death rate.*
- *Texas is 91 deaths per million . . . Texas is 91 deaths per million.*
- *Texas is 91 deaths per million. Texas.*

Non-Examples

- **Fails test 1:** *What you need to know—Well, you need to avoid this.*

- **Fails test 1:** *Texas amazingly has the lowest death rate. New York amazingly has the highest.*
- **Fails test 2:** *What they're not telling you is the death rate. . . . They're not reporting the number of deaths per million.*
- **Fails test 3:** *Oh my God!*
- **Fails test 3:** *Hooray!*

Notes

When a sentence is repeated, the first occurrence of the sentence does not get this label. Only later occurrences do.

Table 39 provides the annotation criteria for Repetition Rhetoric.

Table 39 Repetition Rhetoric annotation criteria

Annotation Criteria	
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.	
Frame Elements	**Claim *X*:** A previous sentence (or phrase)
	Claim *X′*: The current sentence (or phrase) that is a repetition of *X*.
	Test 1: The passage must still make sense and have the same meaning if the repeated phrase/sentence is removed.
	Test 2: The repetition must not add any new information or additional detail.
	Test 3: Vacuous expressives like "*Oh my God!", "Hooray!",* etc., do not count as repetitions.
	Note: The first occurrence of the sentence does not get annotated as Repetition. Only later occurrences do.
Rhetorical Form Example	*[Cats are the best pet]X . . . [Cats are the best]X′.*

2.4.6.2 Rhetoric 9.2: Slogan

Definition

Use of a catchy phrase to support a position or claim.

Examples

- *Cats not rats!*

Table 40 provides the annotation criteria for Slogan Rhetoric.

Table 40 Slogan Rhetoric annotation criteria

Annotation Criteria	
Rule 1: Make sure you can identify each frame element. **Rule 2**: Apply all the **Tests** listed under each frame element. **Rule 3**: If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.	
Frame Elements	**Claim *X*:** Slogan
Rhetorical Form Example	*[Cats not rats!]X*

2.4.7 Manipulating Behavior

Definition

This category includes rhetorical techniques that aim to manipulate the reader's behavior. These are in particular used in advertising, clickbait, and scams, but also appear in various types of misinformation.

2.4.7.1 Rhetoric 10.1: Appeal to Urgency

Definition

Implying or asserting that action on the part of the reader is urgent.

Examples

- *Come buy a cat before time runs out!*

Table 41 provides the annotation criteria for Appeal to Urgency Rhetoric.

Table 41 Appeal to Urgency Rhetoric annotation criteria

Annotation Criteria		
Rule 1: **Rule 2**: **Rule 3**:	Make sure you can identify each frame element. Apply all the **Tests** listed under each frame element. If an element or logical part is allowed to be implicit, it will be marked with **May be Implicit** along with instructions.	
Frame Elements	**Person *A*:**	The experiencer/reader **May be Implicit**: The reader may be unmentioned.
	Event *E*:	Action that is presented as urgent for the reader to do.
	Event *F*:	Feeling of urgency **Common Phrases**: *now, before time runs out, while you still can*
Rhetorical Form Example	*[Come buy a cat]E [before time runs out]F!* Implicit: The [reader]*A* who experiences urgency.	

3. Corpus

We describe the corpus constructed for developing and refining the misinformation annotation schema and detail the annotation procedure as applied to that corpus.

3.1 COVID-19 Corpus

The annotation documents in our corpus are related to the topic of COVID-19, largely from US sources. Each article has one of six focus topics known to be 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 of these topics, there were two opposing stances that were searched for and used in annotations. 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 United States 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 United States 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 manmade 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.

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. The number of articles corresponding to a particular topic and genre are summarized in Table 42. Not all of our documents are full original texts; for example, the scientific journal sources only include the abstracts. All of the annotators were only allowed to annotate the text included in the document, so they were not made aware of any ads, images, or outside information included in their original source.

Table 42 Corpus by topic and genre

Topic	Genre
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 (2) Health care sites (2)
SARS-Cov2 origin	General news (2) Scientific article (2)

3.2 Annotation Procedure

In the first pass of annotation, each individual document, with information on its genre and topic, was presented to three annotators (authors of this report and native English speakers with linguistics training living in the United States) in a separate 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 annotators applied the litmus test given for each layer of annotation and proceeded with giving each sentence a specific label to that layer or "none." This process of reading and labeling sentences was done for three separate layers that included the motivation, fallacy, and rhetoric layers. Annotators were trained for this task by utilizing discussions among one another and applying the Litmus Test, so they would be able to annotate each document with prior knowledge of the misinformation taxonomy given. Once each annotator had completed this task, they would meet again to discuss and decide on the Gold Standard for that document. The Gold Standard is what the annotators each decided on as the right label for each sentence. This annotation process took place over time with annotators updating the taxonomy with more research and suggestions for improvements from the results of their Inter-Annotator Agreement (IAA).

The discussion of these annotations between the three annotators helped to decide on labels to do away with, consolidate, and which ones required too much outside knowledge to be able to train a system to replicate. This fallacy annotation task started with the original five labels from Argotario, which turned into 34 labels with the addition of SemEval's labels in Version 1.0. Then 62 labels were included in Version 2.0 with the addition of sub-labels. And in Version 3.0 there were 69 labels at one point. All of these versions mentioned and their tables are located in the History.

4. 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 data sets focused broadly on the detection and recognition of misinformation, which may have some overlapping categories with our research on fallacies, including the SemEval 2020 annotated data set (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.

In addition to the Argotario corpus, Da San Martino et al. (2019) annotate a corpus for various propaganda techniques, including the annotation of 18 fallacies, of which 10 categories overlap with the fallacies of interest in this paper: Repetition, Appeal to Fear, Flag Waving, Slogan, Appeal to Authority (Irrelevant Authority) Thought-Terminating Cliché, Whataboutism, Reductio ad Hitlerum, Bandwagon, and Straw Man. As in our approach, the authors annotate journal articles as opposed to eliciting or seeking out particular fallacies. As a result, the annotated corpus suffers from an imbalance of fallacies that the authors conclude to be problematic for use of the corpus as training data. This is consistent with our own earlier findings when leveraging a preliminary corpus annotated with just five fallacies as training data (Bonial et al. 2022a) and underscores the need for a carefully curated and balanced training corpus.

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. An overlapping category between their schema and our own is *Hasty Generalization*, for which the authors report the lowest IAA of any of their categories, measured via Cohen's κ, of 0.38. The highest IAA reported is 0.64 for *Appeal to Authority*. The relatively low scores underscore the challenge of this annotation task. The authors explore several models for automatic prediction of the fallacies, including Bidirectional Encoder Representations from Transformers (BERT) and Multimodal Graph Network (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 automatic prediction of *Hasty Generalization*.

Thus, from evaluations documented in related work as well as our own preliminary evaluation of the existing *Argotario* schema, we conclude that while a crowdsourcing approach to misinformation annotation 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. Thus, the schema described in the present work provides rigorously and uniquely defined non-overlapping categories that can be reliably arrived at through a series of annotation decisions in a decision tree.

5. Conclusions and Future Work

The CAMPFIRE annotation schema provides a detailed strategy for identifying features of misinformation that are internal to text, including fallacies and propaganda techniques. Because misinformation often spreads quickly after sudden events and dedicated resources for fact checking do not become available until later, identifying misinformation based on internal features is an important subtask for NLP systems which designed to inform users. At the same time, for internal features to be useful, they must be objective, explainable, and independent of external knowledge. Any annotation of fallacies and propaganda techniques that is subjective or that features too much disagreement between annotators will not be useful as training data or as a trusted strategy for keeping users informed. Annotations that are not explainable will not be useful to users or will be ignored. Annotations that rely on external knowledge will vary between annotators who often have differing external knowledge and will not be predictable at runtime when that external knowledge likely is not available. CAMPFIRE was designed with these constraints in mind and relies on clearly defined annotation tests to evaluate if given fallacy or rhetorical technique is present without relying on external knowledge. Annotated labels in CAMPFIRE are chosen to be explainable and are organized in a taxonomy that makes it easy to explain both what a fallacy or rhetorical technique looks like and why it is misleading.

In future work, we hope to further revise and expand the CAMPFIRE schema to use discourse features to more accurately identify fallacies, to include document-level features such as document misinformation type, and add labels to our fallacy and rhetorical technique taxonomies as we identify new examples in our data. 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. Nonetheless, demonstrating a solid baseline system trained on our data for recognizing misinformation indicators will be a final step in demonstrating the quality of the annotation schema and its replicability for both humans and automatic systems.

6. References

Bonial C, Blodgett A, Hudson T, Lukin SM, Micher J, Summers-Stay D, Sutor P, Voss CR. The search for agreement on logical fallacy annotation of an infodemic. In: Proceedings of the 13th International Conference on Language Resources and Evaluation; 2022a; Marseille, France.

Bonial C, Hudson TA, Blodgett A, Lukin SM, Micher J, Summers-Stay D, Sutor P, Voss CR. You can't quarantine the truth: lessons learned in logical fallacy annotation of an infodemic. DEVCOM Army Research Laboratory (US); 2022b. Report No.: ARL-TR-9343.

Bradd S. Infodemic. World Health Organization; 2021 Nov [2022 Sep]. https://www.who.int/health-topics/infodemic.

Da San Martino G, Yu S, Barrón-Cedeño A, Petrov R, Nakov P. 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); 2019; Hong Kong, China. p. 5636–5646. Association for Computational Linguistics.

Da San Martino G, Barron-Cedeno A, Wachsmuth H, Petrov R, Nakov P. SemEval-2020 task 11: detection of propaganda techniques in news articles. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation; 2020a Dec; Barcelona (online). p. 1377–1414. International Committee for Computational Linguistics.

Da San Martino G, Cresci S, Barrón-Cedeño A, Yu S, Pietro RD, Nakov P. A survey on computational propaganda detection. In: Bessiere C, editor. Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI-20; 2020b. p. 4826–4832. International Joint Conferences on Artificial Intelligence Organization, 7. Survey track.

Habernal I, Hannemann R, Pollak C, Klamm C, Pauli P, Gurevych, I. Argotario: computational argumentation meets serious games. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; 2017.

Sahai S, Balalau O, Horincar R. 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; 2021 Aug; Online. p. 644–657. vol 1; Long Papers. Association for Computational Linguistics.

Zhang AX, Ranganathan A, Metz SE, Appling S, Sehat CM, Gilmore N, Adams NB, Vincent E, Lee J, Robbins M, et al. A structured response to misinformation: defining and annotating credibility indicators in news articles. In: Companion Proceedings of The Web Conference; 2018. p. 603–612.

Appendix A. Schema Application and Annotation Example

A.1 How to Read Annotation Criteria

The criteria for determining when a particular technique is being used are described by a table like the following:

<table>
<tr><th colspan="4">Annotation Criteria</th></tr>
<tr><td></td><td colspan="3">Rule 1: Make sure you can identify each frame element and the premise and conclusion.
Rule 2: Apply all the Tests listed under each frame element.
Rule 3: If an element or logical part is allowed to be implicit, it will be marked with May be Implicit along with instructions.
Rule 4: Sometimes both the premise and conclusion will appear as separate sentences. When that happens, you should annotate whichever one is denoted the “Target” in the logical form.</td></tr>
<tr><td rowspan="3">Frame Elements</td><td colspan="3">Element A: Description</td></tr>
<tr><td colspan="3">May be Implicit: A description of when this element is allowed to be implicit (i.e., implied by the sentence but not stated explicitly).</td></tr>
<tr><td colspan="3">Element B: Description
Test 1: Any rules or requirements</td></tr>
<tr><td rowspan="2">Logical Form (Fallacies only)</td><td>Premise (Target)</td><td>: Description</td><td>Ex: An example premise</td></tr>
<tr><td>Conclusion :</td><td>Description</td><td>Ex: An example conclusion</td></tr>
<tr><td>Rhetorical Form Example (Rhetoric only)</td><td colspan="3">An example sentence with [bracketed spans]B to indicate the presence of a frame element.</td></tr>
</table>

A.2 Misinformation Taxonomy

Fallacy	Rhetoric
1. Deductive Fallacy	1. Negative Emotion
1.1. Begging the Question	1.1. Appeal to Anger
1.2. Black and White Fallacy	1.2. Appeal to Fear
1.3. Association Fallacy	1.2.1. War Metaphor
1.3.1. Appeal to Nature	1.3. Appeal to Sadness/Pity
1.3.2. Appeal to Novelty	2. Positive Emotion
1.3.3. Appeal to Tradition	2.1. Appeal to Optimism
1.4. Thought-Terminating Cliché	2.2. Appeal to Flattery
2. Inductive Fallacy	2.3. Appeal to Loyalty
2.1. Hasty Generalization	2.3.1. Flag Waving
2.2. Appeal to Accident	3. Saliency Bias
2.3. Appeal to Ignorance	3.1. Cliffhanger
2.4. Causal Fallacy	3.2. Dramatization
2.4.1. Correlation–Causation Fallacy	3.3. Vividness
2.4.2. Post Hoc Fallacy	4. Memory Bias
2.4.3. Slippery Slope	4.1. Repetition
3. Abductive Fallacy	4.2. Slogan
3.1. Conspiracy Theory	5. Manipulating Behavior
3.2. Scapegoat	5.1. Appeal to Urgency
4. Testimony Fallacy	
4.1. Bandwagon	
4.2. Irrelevant Authority	
4.3. Sourceless Testimony	
4.4. Appeal to Confidence/Disbelief	
4.4.1. Plain Folks Fallacy	
5. Rebuttal Fallacy	
5.1. Appeal to Conspiracy	
5.2. Appeal to Cover-up	
5.3. Rebuttal by Ad Hominem	
5.3.1. Rebuttal by Tone	
5.3.2. Reductio Ad Hitlerum	
5.4. Straw Man Generalization	
5.5. Two Wrongs Make a Right	
5.5.1. Whataboutism	

A.3 Gold Annotations

The following shows a single document with gold Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably (CAMPFIRE) annotations.

Sentence	Motivation	Fallacy	Rhetoric
Gigantic COVID-19 news, and the thrust of most of the COVID-19 news is to get you eventually to fall in with shutting down the country again.	Distrust, Inform	Straw Man Generalization	None
That's where all of this is headed.	Attention, Distrust	None	None
What you need to know — Well, you need to avoid this.		None	None
You need to not go along with this idea of shutting down the entire economy again, shutting down entire states.	Distrust, Inform	None	None
What they're not telling you is the death rate.	Distrust, Inform	Appeal to Cover-up	None
The death rate is falling.	Inform	None	None
But you don't know that because they are simply reporting this massive increase in cases.	Distrust	Appeal to Cover-up	None
And you're supposed to assume that every case equals a death.	Distrust	Appeal to Conspiracy	None
This is due to media irresponsibility and propaganda as well.	Distrust	Appeal to Conspiracy	None
So record number of cases in Florida set every day.	Attention	None	None
Not in deaths, folks.	Inform	None	None
The number — In fact, you'd be stunned.	Attention	None	Cliffhanger
You probably think that the number-one state for deaths due to coronavirus — What would you say?	Attention	None	Cliffhanger
Probably Florida, right, based on the reporting?	Attention, Distrust	None	Cliffhanger
I mean, 10,000 new cases a day, every new day is a record.	Attention	None	None
There are 11,000 new cases!	Attention	None	Appeal to Fear
Oh, my God.	Attention	None	Appeal to Fear
Shutting down the bars and restaurants in Miami.	Attention	None	Appeal to Fear
Oh, my God!	Attention	None	Appeal to Fear
Shutting down in Palm Beach.	Attention	None	Appeal to Fear
Oh, my God!	Attention	None	Appeal to Fear
You think that death is crazy high in Florida and in Texas.	Attention	None	None
Notice all these states that they’re reporting have these massive increases in cases — Texas, Florida.	Distrust, Inform	None	None
(They don’t tell you about Georgia, by the way.)	Attention, Distrust	Appeal to Cover-up	None

Sentence	Motivation	Fallacy	Rhetoric
Well, I have a little chart here, folks, “COVID-19 Deaths Per Million,” and “the number-one state for deaths -” and this is not a factor of population because this is per million.	Inform	Sourceless Testimony	None
The number-one state where death is occurring to COVID-19’s New Jersey.	Inform	None	None
Wait a minute, New Jersey?	Attention	None	Repetition
I don't even hear about New Jersey on the news.	Attention, Distrust	Appeal to Cover-up	None
New Jersey?	Attention	None	Repetition
They're not even telling me about the number of new cases in New Jersey.	Attention, Distrust	Appeal to Cover-up	None
All I'm hearing about is Florida and Texas.	Attention, Distrust	Appeal to Cover-up	None
Yep.		None	None
It’s 1,718 deaths per million in New Jersey.	Inform	None	None
Do you know what number two is?	Attention	None	Cliffhanger
Number two happens to be Andrew Cuomo’s state, New York, with 1,656 deaths per million.	Inform	None	None
New York?	Attention	None	Repetition
Why, they’re telling me New York is on the downturn.	Attention, Distrust	None	None
They’re telling me New York is getting better.	Attention, Distrust	None	Repetition
They’re telling me no new cases in New York, right?	Attention, Distrust	None	Repetition
Or it’s really trending down.	Attention, Distrust	None	Repetition
All I’m hearing about is Florida.	Attention, Distrust	Appeal to Cover-up	Repetition
All I’m hearing about is Texas.	Attention, Distrust	Appeal to Cover-up	Repetition
Okay, well, let’s find Florida.		None	None
Florida is 172 deaths per million—172?	Attention, Inform	None	Repetition
Well, let’s see.		None	None
The United States average, by the way, the 400.	Inform	None	None
So Florida—Let’s see Texas.		None	None
Texas is 91 deaths per million.	Inform	None	None
Texas.	Attention	None	Repetition
And yet the news you hear every day is all these new cases.	Distrust	None	None
Massive new cases, record numbers of new cases in Florida, in Texas.	Attention	None	Repetition
We gotta shut down the country.	Attention	None	Appeal to Fear
Oh, my God!	Attention	None	Appeal to Fear
It’s getting out of control again.	Attention	None	Appeal to Fear

Sentence	Motivation	Fallacy	Rhetoric
Well, the number of people dying in Texas and Florida — and of course one is too many.	Concession	None	Appeal to Sadness/Pity
Don't misunderstand.	Attention	None	None
But we're nowhere near — In Florida or Texas, we're nowhere near the leading states, which are New York and New Jersey.	Inform	Intuition/Bias Fallacy	None
Following that, by the way, Connecticut, Massachusetts, Rhode Island, DC, Louisiana, Michigan, Illinois.	Diversion	None	None
So you have to get—		None	None
Let's see.		None	None
Ohio is 251 deaths per million.	Inform	None	None
Let's see.		None	None
Let me find Arizona 'cause it's another state they're pushing big.	Distrust	Conspiracy Theory	None
Where is Arizona?		None	None
Arizona may be so low it's not on the list.	Attention	None	None
Arizona is at 250, it looks like, right under Ohio, at 251.	Inform	None	None
But you wouldn't know this, because they're not telling you.	Distrust	Appeal to Cover-up	None
They're not reporting the number of deaths per million.	Distrust	Appeal to Cover-up	None
In other words, they're not reporting the survivability rate.	Distrust	Appeal to Cover-up	None
The answer here is don't mandate closures.	Inform	Correlation–Causation Fallacy	None
Don't mandate social distancing.	Inform	Correlation–Causation Fallacy	None
Don't even mandate mask wearing.	Inform	Correlation–Causation Fallacy	None
Encourage people who are old or who have a compromised immune system to stay quarantined, stay hidden away.	Inform	None	None
Do not go out.	Attention	None	None
But let the young and the healthy go out and live their lives.	Inform	None	Appeal to Joy/Hope
Go ahead and live their lives and spread herd immunity because that's ultimately—'til we get therapeutics or a vaccine, herd immunity is—gonna be the answer to this.	Inform	Correlation–Causation Fallacy	Appeal to Joy/Hope

Appendix B. Schema History

B.1 Version 1.0

False Argument 1. Begging the Question 2. Hasty Generalization 3. Correlation–Causation Fallacy 4. Slippery Slope 5. Appeal to Ignorance 6. Black & White Fallacy 7. Exaggeration/Minimization 8. Causal Oversimplification	Distrust/Discredit 1. Ad Hominem 2. Reductio ad Hitlerum	Emotion 1. Appeal to Fear/Prejudice 2. Appeal to Anger 3. Appeal to Sadness/Pity 4. Appeal to Joy/Hope 5. Appeal to Flattery 6. Flag Waving
False Evidence 1. Irrelevant Authority 2. Bandwagon 3. Hearsay 4. Appeal to Nature 5. Appeal to Novelty 6. Appeal to Tradition 7. Cherry Picking 8. Appeal to Accident	Diversion 1. Red Herring 2. Whataboutism 3. Straw Man 4. Obfuscation 5. Thought-terminating Cliché	Engagement 1. Slogan 2. Repetition 3. Sensationalism 4. Vividness 5. Cliffhanger

Key

Fallacy 🟠 Style/Tone 🔵

B.2 Issue: Using Logical Forms as a Litmus Test

The premise and conclusion can be the same or different sentences. Either the premise or conclusion might be implicit:

- Two wrongs make a right (related to *whataboutism*). Assuming that some wrongful action is justified if the same or similar action was done by someone else. Ex: *Sure, I stole my neighbor's cat, but my other neighbor stole my dog before that!*

  Premise: Person A did X. Conclusion: It is acceptable for person B to do X.

- Slippery Slope. Asserting without evidence that some event will inevitably lead to a more extreme event. Ex: *If we allow pet cats, it's just a matter of time until someone has a pet alligator.*

  Premise: Event A occurs. Conclusion: Extreme event B will occur.

B.3 Version 2.0

False Argument 1. Begging the Question 2. Unsubstantiated Claim 3. Hasty Inference a. Hasty Generalization b. Correlation–Causation Fallacy c. Slippery Slope d. Appeal to Ignorance 4. Reductionism a. Black & White Fallacy b. Exaggeration/Minimization c. Causal Oversimplification	Distrust/Discredit 1. Ad Hominem 2. Reductio ad Hitlerum 3. Appeal to Conspiracy 4. Appeal to Cover-up 5. Appeal to Prejudice	Emotion 1. Appeal to Fear a. Appeal to Urgency b. War Metaphor 2. Appeal to Anger 3. Appeal to Sadness/Pity 4. Appeal to Joy/Hope 5. Appeal to Flattery 6. Appeal to Loyalty a. Flag Waving
False Evidence 1. Irrelevant Testimony a. Irrelevant Authority b. Bandwagon c. Hearsay d. Appeal to Confidence 2. Irrelevant Data a. Appeal to Nature b. Appeal to Novelty c. Appeal to Tradition d. Two Wrongs Make a Right e. False Analogy 3. Misrepresenting Evidence a. Cherry Picking b. Appeal to Accident	Diversion 1. Leading the Witness 2. Labelling 3. Red Herring a. Whataboutism b. Straw Man 4. Halting Discussion a. Obfuscation b. Thought-terminating Cliché	Engagement 1. Slogan 2. Repetition 3. Humor 4. Mockery a. Name Calling 5. Dramatization a. Vividness b. Creating Suspense

Key

Fallacy 🟠 Style/Tone 🔵

Three labels may be annotated for small spans:

- Appeal to Prejudice (S, NP, AP)
- Labeling (NP, AP)
- Name Calling (NP, AP)

B.3.1 Issue: Treating Misinformation Labels as Frames

Reductio Ad Hitlerum

Frame Components:

- Person/Group A : Target of distrust

- Person/Group B : Evil or disliked person/group
    - Common examples: Nazis, fascists, Adolf Hitler, Communists, Socialists
- Claim X : Claim made by A
- Claim X' : Claim made by B
    - Required: X' is perceived as similar to X by the author

	Litmus Test
Premise	**B reports X'**
Conclusion	**A, who reports X, agrees with or is the same as B.**

B.3.2 Issue: How to Deal with Labels that Require External Knowledge?

False Argument 1. Begging the Question 2. Unsubstantiated Claim 3. Hasty Inference a. Hasty Generalization b. Correlation–Causation Fallacy c. Slippery Slope d. Appeal to Ignorance 4. Reductionism a. Black and White Fallacy b. Exaggeration/Minimization c. Causal Oversimplification	Creating Distrust 1. Dismissing Evidence a. Ad Hominem b. Appeal to Conspiracy c. Appeal to Cover-up 2. Conjuring a Villain a. Conjuring a Conspiracy 3. Reductio ad Hitlerum	Emotion 1. Appeal to Fear a. War Metaphor 2. Appeal to Anger 3. Appeal to Sadness/Pity 4. Appeal to Joy/Hope 5. Appeal to Flattery 6. Appeal to Loyalty a. Flag Waving
False Evidence 1. Irrelevant Testimony a. Irrelevant Authority b. Bandwagon c. Hearsay d. Appeal to Confidence 2. Irrelevant Data a. Appeal to Nature b. Appeal to Novelty c. Appeal to Tradition d. False Analogy e. Two Wrongs Make a Right 3. Misrepresenting Evidence a. Cherry Picking b. Appeal to Accident	Diversion 1. Leading the Witness 2. Red Herring a. Whataboutism b. Straw Man 3. Halting Discussion a. Obfuscation b. Thought-terminating Cliché	Engagement 1. Slogan 2. Repetition 3. Appeal to Urgency 4. Appeal to Humor 5. Mockery 6. Dramatization a. Vividness b. Creating Suspense

No External Knowledge	**Requires General Knowledge**	**Requires Direct Fact-Checking**
Appeal to Fear	Straw Man?	False Claim
Black and White Fallacy	Irrelevant Authority?	Exaggeration/Minimization
Slippery Slope		Misrepresented Quote
etc.		False Causal Inference
		Reductionism?
		Causal Oversimplification?
		False Analogy?
		Cherry Picking?

B.4 Version 3.0

Persuasion

1. Deductive Fallacy
   a. Begging the Question
   b. Black and White Fallacy
   c. ~~Reductionism~~
      i. ~~Exaggeration/Minimization~~
2. Inductive Fallacy
   a. Hasty Generalization
   b. Appeal to Ignorance
   c. Slippery Slope
   d. Appeal to Accident
3. ~~Probabilistic Fallacy~~
   a. ~~Cherry Picking~~
4. Causal Fallacy
   a. Correlation–Causation Fallacy
   b. ~~Overlooking Causal Factors~~
5. Testimony Fallacy
   a. Irrelevant Authority
   b. Bandwagon
      i. ?Plain Folks
   c. Hearsay
   d. Appeal to Confidence/Disbelief
6. Intuition/Bias Fallacy
   a. Appeal to Nature
   b. Appeal to Novelty
   c. Appeal to Tradition
   d. ~~False Analogy~~
   e. Two Wrongs Make a Right
7. Abductive Fallacy
   a. Unfalsifiable claim (c.f., Popper's Rule)

Creating Distrust

1. Rebuttal Fallacy
   a. Ad Hominem
   b. Reductio ad Hitlerum
   c. Appeal to Conspiracy
   d. Appeal to Cover-up
   e. Appeal to Hypocrisy
      i. Whataboutism
2. Abductive Fallacy
   a. Conspiracy Theory (c.f., Occam's Razor)
   b. Scapegoat (c.f., Occam's Razor?)
   c. ?Pseudoscientific Hypothesis (c.f., Popper's Rule)
   d. ?Assumption of Malice (c.f., Hanlon's Razor)

Emotion

1. Appeal to Fear
   a. War Metaphor
2. Appeal to Anger
3. Appeal to Sadness/Pity
4. Appeal to Joy/Hope
5. Appeal to Flattery
6. Appeal to Loyalty
   a. Flag Waving

Diversion

1. Diverting Discussion
   a. Red Herring
   b. Straw Man
   c. ~~Leading the Witness~~
2. Avoiding Discussion
   a. Obfuscation
   b. Thought-terminating Cliché

Engagement

1. Slogan
2. Repetition
3. Appeal to Urgency
4. ~~Appeal to Humor~~
5. Mockery
6. Dramatization
   a. Vividness
   b. Creating Suspense

Key

Empirical/Probabilistic ● Causal ● Deductive ● Intuition/Bias ● Testimony ●

Rebuttal ● Abductive ● Discourse Effect ● Emotion/Engagement ●

B.4.1 Issue: Incorporating Inference Types?

1) Deductive
    a. If A logically entails B then $A \Rightarrow B$
2) Inductive and Empirical
    a. Given observations of $A \wedge B$ and none (few) of $A \wedge \neg B$, then $A \Rightarrow B$
3) Causal
    a. If A causes B then $A \Rightarrow B$
    b. (A causes B if A correlates with B and, while controlling for all other factors, enacting $\neg A$ results in $\neg B$)
4) Probabilistic
    a. $Pr(A \wedge B)$ is high and $Pr(A \wedge \neg B)$ is low, then $A \Rightarrow B$
5) Testimony
    a. A reports $B \Rightarrow B$
6) Legal & Social
    a. B is expected by law or social convention $\Rightarrow B$
7) Intuition and Bias
    a. B appeals to someone's intuition and/or bias $\Rightarrow B$

Deductive Fallacy

1) Begging the Question
2) Reductionism
    a. Black and White Fallacy
    b. Exaggeration/Minimization

Inductive Fallacy

1) Hasty Inference
    a. Hasty Generalization
    b. Slippery Slope
    c. Appeal to Ignorance

d. Cherry Picking

e. Appeal to Accident

Causal Fallacy

1) Correlation–Causation Fallacy

2) Overlooking Causal Factors

Testimony Fallacy

1) Irrelevant Testimony

a. Irrelevant Authority

b. Bandwagon

c. Hearsay

d. Appeal to Confidence

Bias/Intuition Fallacy

1) Irrelevant Data

a. Appeal to Nature

b. Appeal to Novelty

c. Appeal to Tradition

d. False Analogy

e. Two Wrongs Make a Right

Misinformation Cue	
Fallacy	**Rhetoric**
1. Deductive Fallacy 1.1. Begging the Question 1.2. Black and White Fallacy 2. Inductive Fallacy 2.1. Hasty Generalization 2.1.1. Slippery Slope 2.2. Appeal to Accident 2.3. Appeal to Ignorance 3. Causal Fallacy 3.1. Correlation–Causation Fallacy 4. Testimony Fallacy 4.1. Irrelevant Authority 4.2. Bandwagon 4.2.1. Plain Folks Fallacy 4.3. Hearsay 4.4. Appeal to Confidence/Disbelief 5. Intuition/Bias Fallacy 5.1. Appeal to Nature 5.2. Appeal to Novelty 5.3. Appeal to Tradition 5.4. Thought-terminating Cliché 5.5. Two Wrongs Make a Right 5.5.1. Whataboutism 6. Abductive Fallacy 6.1. Conspiracy Theory 6.2. Scapegoat 7. Rebuttal Fallacy 7.1. Appeal to Conspiracy 7.2. Appeal to Cover-up 7.3. Rebuttal by Ad Hominem 7.4. Reductio ad Hitlerum 7.5. Straw Man	1. Negative Emotion a. Appeal to Anger b. Appeal to Fear i. War Metaphor c. Appeal to Sadness/Pity 2. Positive Emotion a. Appeal to Joy/Hope b. Appeal to Flattery c. Appeal to Loyalty i. Flag Waving 3. Saliency Bias a. Dramatization b. Vividness c. Appeal to Surprise? 4. Memory Bias a. Repetition b. Slogan 5. Manipulating Behavior a. Appeal to Urgency b. Cliffhanger?
Fallacy: Deductive ● Inductive ● Causal ● Testimony ● Intuition/Bias ● Abductive ● Rebuttal ● **Rhetoric:** Neg Emotion ● Pos Emotion ● Engagement ●	

Motive	
Persuasion Creating Distrust	Engagement Diversion

Appendix C. Symbols and Notation

C.1 Notations

Arrow (⇒) is used to denote credible inferences of all types, including deductive, inductive, abductive, and testimonial inference. X ⇒ Y means that Y can be inferred from X (for a given type of inference).

Definition $(\stackrel{\text{def}}{=})$ is used to offer definitions, particularly when defining different inference types.

Subscripts (word_1, word_2, . . .) are sometimes used in descriptions of annotation criteria for techniques where more than one version is possible. Words with the same subscript correspond to the same version of that technique.

Capital letters

- $A, B, C, \ldots$ denote people and groups of people
- $E, E', F, F', \ldots$ denote events and observables
- $X, X', Y, Y', \ldots$ denote propositions (truth claims)

Brackets ([. . .]A) indicate that a phrase or span corresponds to a frame element A.

List of Symbols, Abbreviations, and Acronyms

ARL	Army Research Laboratory
BERT	Bidirectional Encoder Representations from Transformers
CAMPFIRE	Combined Annotations of Misinformation, Propaganda, and Fallacies Identified Robustly and Explainably
DEVCOM	US Army Combat Capabilities Development Command
MGN	Multimodal Graph Network
NLP	Natural Language Processing
PPI	potentially problematic information

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB
2 (PDF)	DEVCOM ARL FCDD RLA IC C BONIAL C VOSS