Misinfo Reaction Frames:
Reasoning about Readers' Reactions to News Headlines

Saadia Gabriel[♠] **Skyler Hallinan**[♠] **Maarten Sap**[♢♡] **Pemi Nguyen**[♠]
Franziska Roesner[♠] **Eunsol Choi**[♣] **Yejin Choi**[♠♢]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Department of Computer Science, The University of Texas at Austin
[♢]Allen Institute for Artificial Intelligence
[♡]Language Technologies Institute, Carnegie Mellon University
{skgabrie,hallisky,peming,franzi,yejin}@cs.washington.edu,
maartensap@cmu.edu, eunsol@utexas.edu

Abstract

Even to a simple and short news headline, readers *react* in a multitude of ways: cognitively (e.g. inferring the writer's intent), emotionally (e.g. feeling distrust), and behaviorally (e.g. sharing the news with their friends). Such reactions are instantaneous and yet complex, as they rely on factors that go beyond interpreting factual content of news.

We propose **Misinfo Reaction Frames** (MRF), a pragmatic formalism for modeling how readers might react to a news headline. In contrast to categorical schema, our free-text dimensions provide a more nuanced way of understanding intent beyond being benign or malicious. We also introduce a Misinfo Reaction Frames corpus, a crowdsourced dataset of reactions to over 25k news headlines focusing on global crises: the Covid-19 pandemic, climate change, and cancer.

Empirical results confirm that it is indeed possible for neural models to predict the prominent patterns of readers' reactions to previously unseen news headlines. Additionally, our user study shows that displaying machine-generated MRF implications alongside news headlines to readers can increase their trust in real news while decreasing their trust in misinformation. Our work demonstrates the feasibility and importance of pragmatic inferences on news headlines to help enhance AI-guided misinformation detection and mitigation.

1 Introduction

> *Many objects, persons, and experiences in the world are framed in terms of their potential role in supporting, harming, or enhancing people's lives or interests. We can know that this is so if we know how to interpret expressions in which such things are evaluated...*
> - Charles J. Fillmore (1976)

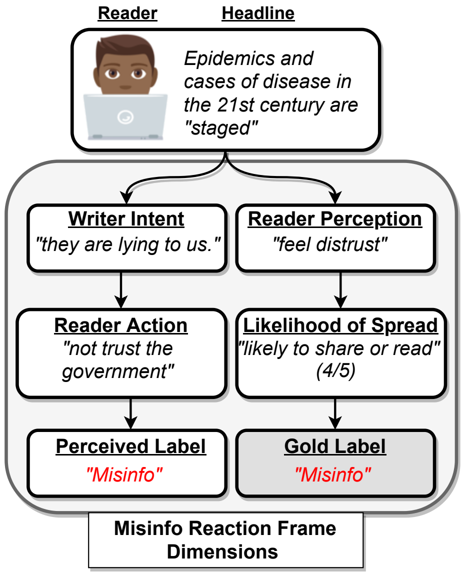


Figure 1: In a binary classification setup, the reaction of the reader is unclear. Here, we use Misinfo Reaction Frames to understand how the reader perceives and reacts to the headline. Our pragmatic frames explain how a health or climate news article is interpreted as reliable or misinformation by readers by incorporating not only linguistic knowledge (e.g. emotions invoked by certain content words), but also knowledge of common social behaviors and domain-specific reasoning. We also include fact-checked labels (gold label).

Effectively predicting how a headline may influence a reader requires knowledge of how readers perceive the intent behind real and fake news. While most prior NLP research on misinformation has focused on fact-checking, preventing spread of misinformation goes beyond determining veracity (Schuster et al., 2020; Ren et al., 2021).

For example, in Figure 1, mistrust in the government may lead readers to share pandemic conspiracy headlines like *"Epidemics and cases of disease in the 21st century are "staged""* even if they suspect it is misinformation. The widespread circulation of misinformation can have serious neg-

☞News Headline	✍Writer's Intent	✍Reader Reaction	Spread	Real News? (GPT-2 / T5 / Gold)
How COVID is Affecting U.S. Food Supply Chain	**Human**: "the pandemic is interrupting the flow of groceries to consumers" **GPT-2**: "food supplies are being affected by covid" **T5**: "food supply chain is affected by covid"	**Human**: "want to know how their groceries will get to them" **GPT-2**: "want to learn more" **T5**:"want to find out more information"	4.0	✓/ ✓/ ✓
Thai police arrested a cat for disobeying the curfew order.	**Human**: "governments are ludicrous and obtuse." **GPT-2**: "animals can be dangerous" **T5**: "lockdowns are enforced in thailand"	**Human**: "feel disbelief" **GPT-2**: "feel worried" **T5**: "feel shocked"	1.0	✗/ ✗/ ✗
Perspective — I'm a black climate expert. Racism derails our efforts to save the planet.	**Human**: "since climate change will likely affect poorer nations, rich societies are not motivated to help" **GPT-2**: "racism is bad" **T5**: "racism is a problem in society"	**Human:** "want to improve their own behavior towards others" **GPT-2:** "want to learn more" **T5**: "want to take action"	3.0	✓/ ✗/ ✓

Table 1: Example instances in MRF corpus along with generations from reaction inference models fine-tuned on the corpus. We show the predicted writer intent, reader reactions (either a perception or action), and the human-annotated likelihood of the headline being shared or read (Spread). The last column (Real News?) shows model-predicted and gold label on whether the headline belongs to a real news or misinformation source. Our task introduces a new challenge of understanding how news impacts readers. As shown by the examples, large-scale pretrained models (GPT-2, T5) miss nuances present in perceptions of informed readers even when they correctly predict whether the headline is from real news or not.

ative repercussions on readers — it can reinforce sociopolitical divisions like anti-Asian hate (Vidgen et al., 2020; Abilov et al., 2021), worsen public health risks (Ghenai and Mejova, 2018), and undermine efforts to educate the public about global crises (Ding et al., 2011).

We introduce **Misinfo Reaction Frames** (MRF), a pragmatic formalism to reason about the effect of news headlines on readers. Inspired by Frame semantics (Fillmore, 1976), our frames distill the pragmatic implications of a news headline in a structured manner. We capture free-text explanations of readers reactions and perceived author intent, as well as categorical estimates of veracity and likelihood of spread (Table 2). We use our new formalism to collect the MRF corpus, a dataset of 202.3k news headline/annotated dimension pairs (69.8k unique implications for 25.1k news headlines) from Covid-19, climate and cancer news.

We train reaction inference models to predict MRF dimensions from headlines. As shown by Table 1, reaction inference models can correctly label the veracity of headlines (85% F1) and infer commonsense knowledge like *"a cat being arrested for disobeying curfew $\Longrightarrow$ lockdowns are enforced."* However, models struggle with more nuanced implications *"a cat arrested for disobeying curfew $\Longrightarrow$ government incompetence."* We test generalization of reaction frame inference on a new cancer domain and achieve 86% F1 by finetuning our MRF model on 574 annotated examples.

To showcase the usefulness of the MRF framework in user-facing interventions, we investigate the effect of MRF explanations on reader trust in headlines. Notably, in a user study our results show that machine-generated MRF inferences affect readers' trust in headlines and for the best model there is a statistically significant correlation (Pearson's r=0.24, p=0.018) with labels of trustworthiness (§5.3).

Our framework and corpus highlight the need for reasoning about the pragmatic *implications* of news headlines with respect to reader reactions to help combat the spread of misinformation. We publicly release the MRF corpus and trained models to enable further work (`https://github.com/skgabriel/mrf-modeling`).[1] We explore promising future directions (and limitations) in (§6).

[1]The full data annotation setup can be found here: `https://misinfo-belief.github.io/`, for use in extending reaction frames to other news domains.

Dimension	Type	Description	Example
Writer Intent	free-text	A writer intent implication captures **the readers' interpretation of what the writer is implying.**	"some masks are better than others."
Reader Perception	free-text	A reader perception implication describes how readers would *feel* in response to a headline. These inferences include **emotional reactions** and **observations**.	"feeling angry.", "feeling that the event described in the headline would trouble most people."
Reader Action	free-text	A reader action implication captures what readers would *do* in response to a headline. These describe **actions**.	"buy a mask."
Likelihood of Spread	ordinal	To take into account variability in impact of misinformation due to low or high appeal to readers, we use a 1-5 Likert (Likert, 1932) scale to measure the **likelihood of an article being shared or read**. Categories are {*Very Likely*, *Likely*, *Neutral*, *Unlikely*, *Very Unlikely*}.	4/5
Perceived Label	binary	We elicit the perceived label (real/misinfo) of a headline, i.e. **whether it appears to be misinformation or real news to readers**.	real
Gold Label	binary	We include the **original ground-truth headline label** (real/misinfo) that was verified by fact-checkers.	misinfo

Table 2: A description of misinformation reaction frame dimensions.

2 Misinfo Reaction Frames

Motivation for Our Formalism In contrast to prior work on misinformation detection (Ott et al., 2011; Rubin et al., 2016; Rashkin et al., 2017; Wang, 2017; Hou et al., 2019; Volkova et al., 2017; Jiang and Wilson, 2018) which mostly focuses on linguistic or social media-derived features, we focus on the potential impact of a news headline by modeling readers' reactions. This approach is to better understand how misinformation can be countered, as it has been shown that interventions from AI agents are better at influencing readers than strangers (Kulkarni and Chi, 2013).

In order to model impact, we build upon prior work that aims to describe the rich interactions involved in human communication, including semantic frames (Fillmore, 1976), the encoder-decoder theory of media (Hall, 1973)[2], Grice's conversational maxims (Grice, 1975) and the rational speech act model (Goodman and Frank, 2016)[3]. By describing these interactions with free-text implications invoked by a news headline, we also follow from prior work on pragmatic frames of connotation and social biases (Speer and Havasi, 2012; Rashkin et al., 2018; Sap et al., 2019, 2020; Forbes et al., 2020).

While approaches like rational speech acts model both a pragmatic speaker and listener, we take a **reader-centric** approach to interpreting "intent" of a news headline given that the writer's intent is challenging to recover in the dynamic environment of social media news sharing (Starbird et al., 2019). By bridging communication theory, data annotation schema and predictive modeling, we define a concrete framework for understanding the impact of a news headline on a reader.

Defining the Frame Structure Table 1 shows real and misinformation news examples from our dataset with headlines obtained from sources described in §3.1. We pair these headline examples with generated reaction frame annotations from the MRF corpus. Each reaction frame contains the dimensions in Table 2.

We elicit annotations based on a *news headline*, which summarizes the main message of an article. We explain this further in §3.1. An example headline is "*Covid-19 may strike more cats than believed.*" To simplify the task for annotators and ground implications in real-world concerns, we define these implications as relating to one of 7 common themes (e.g. technology or government

[2] This theory proposes that before an event is communicated, a narrative discourse encoding the objectives of the writer is generated.

[3] Here pragmatic interpretation is framed as a probabilistic reasoning problem.

entities) appearing in Covid and climate news.[4] We list all the themes in Table 3, with some themes being shared between topics.

3 Misinfo Reaction Frames Corpus

To construct a corpus for studying reader reactions to news headlines, we obtain 69,885 news implications (See §3.1) by eliciting annotations for 25,164 news headlines (11,757 Covid related articles, 12,733 climate headlines and 674 cancer headlines). There are two stages for collecting the corpus - (1) news data collection and (2) crowd-sourced annotation.

3.1 News Data Collection

A number of definitions have been proposed for labeling news articles based on reliability. To scope our task, we focus on false news that may be unintentionally spread (misinformation). This differs from disinformation, which assumes a malicious intent or desire to manipulate (Fallis, 2014). We examine reliable and unreliable headline extracted from two domains with widespread misinformation: Covid-19 (Hossain et al., 2020) and climate change (Lett, 2017). We additionally test on cancer news (Cui et al., 2020) to measure out-of-domain performance.

Climate Change Dataset We retrieve both trustworthy and misinformation headlines related to climate change from NELA-GT-2018-2020 (Gruppi et al., 2020; Norregaard et al., 2019), a dataset of news articles from 519 sources. Each source in this dataset is labeled with a 3-way trustworthy score (reliable / sometimes reliable / unreliable). We discard articles from "sometimes reliable" sources since the most appropriate label under a binary labeling scheme is unclear. To identify headlines related to climate change, we use keyword filtering.[5] We also use claims from the SciDCC dataset (Mishra and Mittal, 2021), which consists of 11k real news articles from ScienceDaily,[6] and Climate-FEVER (Diggelmann et al., 2020), which consists of more than 1,500 true and false climate claims from Wikipedia. We extract claims with either supported or refuted labels in the original dataset.[7]

Theme	Climate	Covid
Climate Statistics	✓	
Natural Disasters	✓	
Entertainment	✓	
Ideology	✓	
Disease Transmission		✓
Disease Statistics		✓
Health Treatments		✓
Protective Gear		✓
Government Entities	✓	✓
Society	✓	✓
Technology	✓	✓

Table 3: Themes present in articles by each news topic. Some are covered by both climate and Covid domains, while others are domain specific.

Statistic	Train	Dev.	Test	Cancer
Headlines	19,897	2,460	2,133	674
Unique Intents	38,172	4,867	4,388	1,232
Unique Percept.	2,609	538	421	174
Unique Actions	15,036	2,176	1,739	704
Total Pairs	159,564	19,700	17,890	5,227

Table 4: Dataset-level breakdown of statistics for MRF corpus.

Covid-19 Dataset For trustworthy news regarding Covid-19, we use the CoAID dataset (Cui and Lee, 2020) and a Covid-19 related subset of NELA-GT-2020 (Gruppi et al., 2020). CoAID contains 3,565 news headlines from reliable sources. These headlines contain Covid-19 specific keywords and are scraped from nine trustworthy outlets (e.g. the World Health Organization).

For unreliable news (misinformation), we use The CoronaVirusFacts/DatosCoronaVirus Alliance Database, a dataset of over 10,000 mostly false claims related to Covid-19 and the ESOC Covid-19 Misinformation Dataset, which consists of over 200 additional URLs for (mis/dis)information examples.[8][9] These claims originate from social media posts, manipulated media, and news articles, that have been manually reviewed and summarized by fact-checkers.

Cancer Dataset We construct an evaluation set for testing out-of-domain performance using cancer real and misinformation headlines from the

[4] We use a subset of the data (approx. 200 examples per news topic) to manually identify themes. Note that themes are not disjoint and a news article may capture aspects of multiple themes.

[5] We kept any article headline that contained at least one of "environment," "climate," "greenhouse gas," or "carbon tax." We remove noisy examples obtained using these keywords with manual cleaning.

[6] https://www.sciencedaily.com/

[7] The data also includes some claims for which there is not enough info to infer a label. We discard these claims.

[8] https://www.poynter.org

[9] esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset

DETERRENT dataset (Cui et al., 2020), consisting of 4.6k real news and 1.4k fake news articles.

3.2 Annotation Process

In this section we outline the structured annotation interface used to collect the dataset. Statistics for the full dataset are provided in Table 4.

Annotation Task Interface We use the Amazon Mechanical Turk (MTurk) crowdsourcing platform.[10] We provide Figure 2 in the Appendix to show the layout of our annotation task. For ease of readability during annotation, we present a headline summarizing the article to annotators, rather than the full text of the article. Annotators then rate veracity and likelihood of spread based on the headline, as well as providing free-text responses for writer intent, reader perception and reader action.[11] We structure the annotation framework around the themes described in §2.

Quality Control We use a three-stage annotation process for ensuring quality control. In the initial pilot, we select a pool of pre-qualified workers by restricting to workers located in the US who have had at least 99% of their *human intelligence tasks* (hits) approved and have had at least 5000 hits approved. We approved workers who consistently submitted high-quality annotations for the second stage of our data annotation, in which we assessed the ability of workers to discern between misinformation and real news. We removed workers whose accuracy at predicting the label (real/misinfo) of news headlines fell below 70%. Our final pool consists of 80 workers who submitted at least three annotations during the pilot tasks. We achieve pairwise agreement of 79% on the label predicted by annotators during stage 3, which is comparable to prior work on Covid misinformation (Hossain et al., 2020). To account for chance agreement, we also measure Cohen's Kappa $\kappa = .51$, which is considered "moderate" agreement. Additional quality control measures were taken as part of our extensive annotation post-processing. For details, see Appendix A.2.

Annotator Demographics We provided an optional demographic survey to MTurk workers during annotation. Of the 63 annotators who reported ethnicity, 82.54% identified as White, 9.52% as Black/African-American, 6.35% as Asian/Pacific Islander, and 1.59% as Hispanic/Latino. For self-identified gender, 59% were male and 41% were female. Annotators were generally well-educated, with 74% reporting having a professional degree, college-level degree or higher. Most annotators were between the ages of 25 and 54 (88%). We also asked annotators for their preferred news sources. New York Times, CNN, Twitter, Washington Post, NPR, Reddit, Reuters, BBC, YouTube and Facebook were reported as the 10 most common news sources.

4 Modeling Reaction Frames

We test the ability of large-scale language models to predict Misinfo Reaction Frames. For free-text inferences (e.g. writer intent, reader perception), we use generative language models, specifically T5 encoder-decoder (Raffel et al., 2020) and GPT-2 decoder-only models (Radford et al., 2019). For categorical inferences (e.g. the gold label), we use either generative models or BERT-based discriminative models (Devlin et al., 2019). We compare neural models to a simple retrieval baseline (**BERT-NN**) where we use gold implications aligned with the most similar headline from the training set.[12]

4.1 Controlled Generation

For generative models, we use the following input sequence

$$x = h_1 \ldots h_T || s_d || s_t,$$

where h is a headline of length T tokens, $s_t \in \{[covid], [climate]\}$ is a special topic control token, and s_d is a special dimension control token representing one of six reaction frame dimensions. Here $||$ represents concatenation. The output is a short sequence representing the predicted inference (e.g. "*to protest*" for reader action, "*misinfo*" for the gold label). For GPT-2 models we also append the gold output inference $y = g_1 \ldots g_N$ during training, where N is the length of the inference.

Inference We predict each token of the output inference starting from the topic token s_t until the $[eos]$ special token is generated. In the case of data with unknown topic labels, this allows us to jointly predict the topic label and output inference. We decode using beam search, since generations by beam search are known to be less diverse but more

[10] `https://www.mturk.com/`

[11] These news events are either article headlines or claims.

[12] Similarity is measured between headlines embedded with MiniLM, a distilled transformer model (Wang et al., 2020). We use the Sentence-BERT package (Reimers and Gurevych, 2019).

factually aligned with the context (Massarelli et al., 2020).

4.2 Classification

For discriminative models, we use the following input sequence

$$x = [CLS]h_1 \,...\, h_T[SEP],$$

where [CLS] and [SEP] are model-specific special tokens. The output is a categorical inference.

4.3 Training

All our models are optimized using cross-entropy loss, where generally for a sequence of tokens t

$$CE(t) = -\frac{1}{|t|}\sum_{i=1}^{|t|} logP_\theta(t_i|t_1, ..., t_{i-1}).$$

Here P_θ is the probability given a particular language model θ. Since GPT-2 does not explicitly distinguish between the input and output (target) sequence during training, we take the loss with respect to the full sequence. For T5 we take the loss with respect only to the output.

4.4 Masked Fine-Tuning

To improve generalization of MRF models, we use an additional masked fine-tuning step. We first train a language model θ on a set of Covid-19 training examples D_{covid} and climate training examples $D_{climate}$. Then we use the Textrank algorithm (Mihalcea and Tarau, 2004) to find salient keyphrases in D_{covid} and $D_{climate}$, which we term k_{covid} and $k_{climate}$ respectively. We determine domain-specific keyphrases by looking at the complement of $k_{covid} \cap k_{climate}$

$$k'_{covid} = k_{covid} \setminus k_{covid} \cap k_{climate}$$
$$k'_{climate} = k_{climate} \setminus k_{covid} \cap k_{climate},$$

and only keep the top 100 keyphrases for each domain. We mask out these keyphrases in the training examples from D_{covid} and $D_{climate}$ by replacing them with a $< mask >$ token. Then we continue training by fine-tuning on the masked examples. A similar approach has been shown to improve generalization and reduce shortcutting of reasoning in models for event detection (Liu et al., 2020).

5 Experiments

In this section, we evaluate the effectiveness of our proposed framework at predicting likely reactions, countering misinformation and detecting misinformation. We first describe setup for experiments (§5.1), as well as evaluation metrics for classification and generation experiments using our corpus (§5.2.1,§5.2.2). We also show the performance of large-scale language models on the task of generating reaction frames (§5.3) and provide results for classification of news headlines (§5.4).

5.1 Setup

We determine the test split according to date to reduce topical and news event overlap between train and test sets.[13] We use the HuggingFace Transformers library (Wolf et al., 2020). Hyperparameters are provided in Appendix A.3.

5.2 Evaluation Metrics

We compare reaction inference systems using common automatic metrics. We also use human evaluation to assess quality and potential use of generated writer intent inferences.

5.2.1 Automatic Metrics

These metrics include the BLEU (-4) ngram overlap metric (Papineni et al., 2002) and BERTScore (Zhang et al., 2020), a model-based metric for measuring semantic similarity between generated inferences and references. For classification we report macro-averaged precision, recall and F1 scores.[14][15] We use publicly available implementations for all metrics (nltk[16] for BLEU).

5.2.2 Human Evaluation

For human evaluation, we assess generated inferences using the same pool of qualified workers who annotated the original data. We randomly sample model-generated "writer's intent" implications from T5 models and GPT-2 large over 196 headlines where generated implications were unique for each model type.[17] We elicit 3 unique judgements per headline. Implications are templated in the

[13] We use news articles from 2021 and the last two months of 2020 for the test set. We ensure there is no exact overlap between data splits.

[14] We compute these using scikit-learn: https://scikit-learn.org/stable/index.html

[15] For measuring likelihood of spread, predicted and averaged values are rounded to the nearest integer.

[16] https://www.nltk.org/

[17] 98 misinfo and 98 real headlines in the dev. set

	Model	Writer Intent BLEU-4 ↑	Writer Intent BERTScore ↑	Reader Perception BLEU-4 ↑	Reader Perception BERTScore ↑	Reader Action BLEU-4 ↑	Reader Action BERTScore ↑
dev.	BERT-NN	31.45	86.29	35.69	91.04	45.47	84.76
	T5-base	51.48	88.03	31.98	92.87	53.55	85.27
	T5-large	51.30	**88.16**	32.82	**92.94**	57.29	**85.34**
	GPT-2 (small)	**60.68**	87.35	**37.22**	92.21	54.20	84.83
	GPT-2 (large)	54.94	87.74	32.35	92.84	**57.84**	85.00
test	BERT-NN	34.46	86.35	**37.09**	90.84	46.57	84.78
	T5-base	50.63	87.78	32.18	**93.32**	57.37	85.60
	T5-large	50.86	**87.94**	32.89	93.29	**62.10**	**85.88**
	GPT-2 (large)	**60.51**	87.73	34.18	92.51	59.57	85.53

Table 5: Automatic modeling results (generation task). For this table and the following tables, we highlight the best-performing model(s) in **bold**.

Model	Quality (1-5)	Influence on Trust +Trust (%)	Influence on Trust -Trust (%)	Influence on Trust Corr w/ Label (all gens)	Influence on Trust Corr w/ Label (quality ≥ 3)	Socially Acceptable (%)
T5-base	3.61	8.33	7.82	**0.24***	**0.30***	**75.30**
T5-large	**3.74**	7.73	9.76	-0.03	0.09	74.66
GPT-2 (large)	3.46	**9.70**	**13.10**	-0.04	0.10	74.66

Table 6: Human evaluation results (generation task). Cells marked by "*" are statistically significant for $p < .05$.

form *"The writer is implying that [implication]"* for ease of readability.

Overall Quality We ask the annotators to assess the overall quality of generated implications on a 1-5 Likert scale (i.e. whether they are coherent and relevant to the headline without directly copying).

Influence on Trust We measure whether generated implications impact readers' perception of news reliability by asking annotators whether a generated implication makes them perceive the news headline as more (+) or less (-) trustworthy.

Perceived Sociopolitical Acceptability We ask annotators to rate their perception of the beliefs invoked by an implication in terms of whether they represent a majority (mainstream) or minority (fringe) viewpoint.[18]

A/B Testing For A/B testing, annotators are initially shown the headline with the generated implication hidden. We ask annotators to rate trustworthiness of headlines on a 1-5 Likert scale, with 1 being clearly misinformation and 5 being clearly real news. After providing this rating, we reveal the generated implication to annotators and have them rate the headline again on the same scale. Annotators were not told whether or not implications were machine-generated, and we advised annotators to mark generated implications that were copies of the headlines as low quality.

5.3 Generating Reaction Frames

The automatic evaluation results of our generation task are provided in Table 5.

Results We found that the T5-large model was rated as having slightly higher quality generations than the other model variants (Table 6). Most model generations were rated as being *"socially acceptable"*. However in as many as 25.34% of judgements, generations were found to be not socially acceptable.

Interestingly, all models were rated capable of influencing readers to trust or distrust headlines, but effectiveness is dependent on the quality of the generated implication. In particular for T5-base, we found a consistent correlation between the actual label and shifts in trustworthiness scores before and after annotators see the generated writer's intent. Annotators reported that writer intents made real news appear more trustworthy and misinformation less trustworthy.[19]

5.4 Detecting Misinformation

To test if we can detect misinformation using propagandistic content like *loaded or provocative lan-*

[18] We refer to "minority" viewpoint broadly in terms of less frequently adopted or extreme social beliefs, rather than in terms of viewpoints held by historically marginalized groups.

[19] While for most models the trend is a decrease in trust for both real news and misinformation, for the T5-base model there is a statistically significant correlation of $Pearson'sr = .24$ showing shifts in trust align with gold labels.

	Model	Spread P ↑	Spread R ↑	Spread F1 ↑	Reader P ↑	Reader R ↑	Reader F1 ↑	Gold P ↑	Gold R ↑	Gold F1 ↑
dev.	Majority Baseline	7.11	20.00	10.49	29.61	50.00	37.20	26.32	50.00	34.49
	T5-base	**29.92**	27.63	22.77	81.43	76.79	77.72	87.11	87.17	87.13
	T5-large	29.66	**30.08**	**29.04**	82.60	**78.13**	**79.04**	88.21	88.06	88.12
	GPT-2 (small)	26.86	23.76	22.38	78.83	77.29	77.80	84.17	83.75	83.86
	GPT-2 (large)	31.76	28.96	27.59	**82.62**	77.73	78.73	90.33	88.76	89.01
	Prop-BERT	-	-	-	-	-	-	51.82	51.09	46.43
	BERT-large	-	-	-	-	-	-	89.50	89.13	89.24
	Covid-BERT	-	-	-	-	-	-	**90.79**	**90.50**	**90.60**
test	Majority Baseline	7.78	20.00	11.20	27.00	50.00	35.07	31.41	50.00	38.58
	T5-base	31.75	27.02	20.59	85.01	82.55	82.91	80.02	81.16	80.43
	T5-large	31.69	31.98	**30.60**	86.76	84.57	**84.95**	80.75	82.35	81.20
	GPT-2 (large)	34.19	27.58	18.41	83.24	83.24	82.70	80.93	82.05	81.35
	Prop-BERT	-	-	-	-	-	-	48.83	49.26	38.79
	BERT-large	-	-	-	-	-	-	79.45	81.20	79.80
	Covid-BERT	-	-	-	-	-	-	**84.83**	**86.97**	**85.26**
cancer (unsup.)	Prop-BERT	-	-	-	-	-	-	**72.60**	65.00	61.78
	BERT-large	-	-	-	-	-	-	23.12	43.00	30.07
	Covid-BERT	-	-	-	-	-	-	67.87	61.00	56.85
	GPT-2 (large)	**27.24**	23.55	10.95	**64.38**	59.21	54.43	59.16	53.00	43.50
	T5-large	21.87	**24.95**	21.12	62.08	**61.62**	**61.44**	41.13	48.00	35.52
	GPT-2 (large) + masked	22.93	23.94	**21.78**	60.06	55.69	51.00	66.03	**66.00**	**65.99**
	T5-large + masked	21.38	22.79	19.57	54.84	54.41	53.66	65.26	55.00	45.91
cancer (sup.)	GPT-2 (large) + sup	30.32	31.03	27.38	66.97	66.83	66.84	87.13	87.00	86.99
	T5-large + sup	12.17	21.67	10.51	75.30	67.95	66.15	86.00	86.00	86.00

Table 7: Automatic modeling results (classification task). For the unsupervised cancer setting (unsup.), all models are trained on covid/climate data only or another news dataset (Prop-BERT). For the supervised setting (sup.), we fine-tune on 574 cancer news examples.

guage (e.g. "Covid-19 vaccines may be *the worst threat we face*"), we use a pre-trained BERT propaganda detector (Da San Martino et al., 2019) which we denote here as (**Prop-BERT**).[20] For our zero-shot setting, we classify a news event as real if it is not associated with any propaganda techniques and misinformation otherwise. As shown by Table 7, F1 results are considerably lower than task-specific models. This is likely due to the fact both real and misinformation news uses propaganda techniques.

Neural misinformation detection models are able to outperform humans at identifying misinformation (achieving a max F1 of 85.26 compared to human performance F1 of 75.21[21]), but this is still a nontrivial task for large-scale models. When we use **Covid-BERT** (Müller et al., 2020), a variant of BERT pretrained on 160M Covid-related tweets, we see an improvement of 5.46% over BERT without domain-specific pretraining (Table 7). This indicates greater access to domain-specific data helps in misinformation detection, even if the veracity of claims stated in the data is unknown.

Performance on Out-of-Domain Data We test the ability of reaction frames to generalize using 100 cancer-related real and misinformation health news headlines (Cui et al., 2020), see Table 7. For the misinformation detection task, we evaluate gold F1 using the **Prop-BERT** zero-shot model, MRF-finetuned BERT-large, **Covid-BERT**, T5-large and GPT-2 large models. We observe that after one epoch of re-training, masked fine-tuning substantially boosts unsupervised performance of generative MRF models (**GPT-2 large + masked** and **T5-large + masked**), making them more robust than BERT variants. We compare this performance against the T5-large and GPT-2 large model fine-tuned on only 574 cancer examples (**GPT-2 large + sup** and **T5-large + sup**), and observe that this leads to a performance increase of up to *43.49%*, achieving similar F1 performance to our domains with full data supervision.

6 Future Directions and Limitations of Reaction Frames

Our framework presents new opportunities for studying perceived intent and impact of misinfor-

[20]The model predicts if any of 18 known propaganda techniques are used to describe a news event. See the paper for the full list.

[21]We count disagreements as being labeled misinformation here, discarding disagreements leads to F1 of 74.97.

mation, which may also aid in countering and detecting misinformation.

We can estimate content virality. Given the user-annotated labels for likelihood of reading or sharing, we can estimate whether the information in the associated article is likely to propagate.

We can analyze the underlying intents behind headlines. Using annotated writer intents, we can determine common themes and perceived intentions in misinformation headlines across domains (e.g. mistrust of vaccination across medical domains). Given the performance of predictive models highlighted by Tables 5 and 6, we can also extend this analysis to unseen headlines.

We can categorize headlines by severity of likely outcomes. False headlines that explicitly incite violence, or otherwise encourage actions that lead to psychological or physical harm (e.g. not vaccinating) may be deemed more malicious than false headlines with more benign consequences (e.g. some examples of satire). Future work may explore categorizing severity of headlines based on potential harms resulting from implications.

Perceived labels can help us understand which headlines may fool readers. We can use these labels to determine which types of misinformation headlines appear most like real news to generally knowledgeable readers. These may also help in designing misinformation countering systems and better adversarial examples to improve robustness of misinformation detection models.

We can generate counter-narratives to misinformation. Our results indicate it is possible to generate effective explanations for the intent of headlines that discourage trust in misinformation (Section 5.3), see Appendix A.5 for examples. We encourage future work that further improves performance of these models (e.g. through integration of domain knowledge).

Limitations. Given these future directions, we also consider key limitations which must be addressed if we move beyond viewing Misinfo Reaction Frames as a proof-of-concept and use the dataset as part of a large-scale system for evaluating or countering misinformation.

Since we focus on news headlines, the context is limited. The intent of a headline may be different from the actual intent of the corresponding article, especially in the case of clickbait. We find headlines to be suitable as online readers often share headlines without clicking on them (Gabielkov et al., 2016), however future work may explore extending reaction frames to full news articles.

There is also annotator and model bias. Readers involved in our data curation and human evaluation studies are "*generally knowledgeable*," as proved by their ability to discern misinformation from real news. We see this bias as a potential strength as it allows us to find ways to counter misinformation in cases where readers are well-informed but still believe false information. However, annotators may have undesirable political or social biases. In such cases, gender bias may lead an annotator to assume that a politician mentioned in a headline is male or to dismiss inequality concerns raised by a scientist belonging to a minority group as "playing the race card." These biases can also appear in pre-training data, leading to model bias.[22] Subjectivity in annotation is a point of discussion in many pragmatic-oriented tasks, e.g. social norm prediction (Jiang et al., 2021) and toxicity detection (Halevy et al., 2021; Sap et al., 2021). We encourage conscious efforts to recruit diverse pools of annotators so multiple perspectives are considered, and future work on modeling reaction frames can consider learning algorithms that mitigate harmful effects of biases, depending on use case (Khalifa et al., 2021; Gordon et al., 2022).

Lastly, we only consider English-language news and annotate with workers based in the US. It may be that news headlines would be interpreted differently in other languages and cultures.

7 Conclusion

We introduced Misinfo Reaction Frames, a pragmatic formalism for understanding reader perception of news reliability. We show that machine-generated reaction frames can change perceptions of readers, and while large-scale language models are able to discern between real news and misinformation, there is still room for future work. Generated reaction frames can potentially be used in a number of downstream applications, including better understanding of event causality, empathetic response generation and as counter-narratives.

[22]Removing these examples from data curation or trying to control for "annotator neutrality" does not erase the causes that lead to the existence of these biases. The fact that harmful biases can manifest in the viewpoints of informed readers speaks to the pervasiveness of certain stereotypes.

8 Ethical Considerations

There is a risk of frame-based machine-generated reader interpretations being misused to produce more persuasive misinformation. However, understanding the way in which readers perceive and react to news is critical in determining what kinds of misinformation pose the greatest threat and how to counteract its effects. Furthermore, while transformer models have contributed to much of the recent algorithmic progress in NLP research and are the most powerful computational models available to us, work has highlighted shortcomings in their performance on domain-specific text (Moradi et al., 2021) and noted that these models can easily detect their own machine-generated misinformation (Zellers et al., 2019). Therefore, we do not see this potential dual-use case as an imminent threat, but urge implementation of systemic changes that would discourage such an outcome in the future - e.g. regulation that would lead to required safety and fairness measures *before* large-scale systems are deployed in the wild (European Commission, 2021).

We emphasize that annotations may reflect *perceptions* and *beliefs* of annotators, rather than universal truths (Britt et al., 2019). Especially considering demographic homogeneity of online crowdsource workers, we urge caution in generalizing beliefs or taking beliefs held in certain social/cultural contexts to be factual knowledge. We obtained an Institutional Review Board (IRB) exemption for annotation work, and ensured annotators were fairly paid given time estimations.

Broader impact. The rapid dissemination of information online has led to an increasing problem of falsified or misleading news spread on social media like Twitter, Reddit and Facebook (Vosoughi et al., 2018; Geeng et al., 2020). We specifically designed the Misinfo Reaction Frames formalism to allow us to identify and predict *high-impact* misinformation that is more likely to spread. This can allow for future research on factors that make misinformation particularly dangerous, as well as systems that are more effective at mitigating spread.

Acknowledgements

The authors thank members of the DARPA SemaFor program, UW NLP, the UW CSE 599 social computing class and Amy X. Zhang for helpful discussions, as well as the anonymous reviewers and Akari Asai for comments on the draft. This research is supported in part by NSF (IIS-1714566), NSF (2041894), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), DARPA SemaFor program, and Allen Institute for AI.

References

Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. Voterfraud2020: a multimodal dataset of election fraud claims on twitter. In *Proceedings of AAAI 2021*.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

M. Britt, J. Rouet, Dylan Blaum, and K. Millis. 2019. A reasoned approach to dealing with fake news. *Policy Insights from the Behavioral and Brain Sciences*, 6:101 – 94.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *ArXiv*, abs/2006.00885.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. Deterrent: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 492–502, New York, NY, USA. Association for Computing Machinery.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

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

Arkin Dharawat, Ismini Lourentzou, Alex Morales, and ChengXiang Zhai. 2020. Drink bleach or do what now? covid-hera: A dataset for risk-informed health decision making in the presence of covid19 misinformation.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *Tackling Climate Change with Machine Learning Workshop at NeurIPS 2020*.

Ding Ding, Edward W. Maibach, Xiaoquan Zhao, Connie Roser-Renouf, and Anthony Leiserowitz. 2011. Support for climate policy and societal action are linked to perceptions about scientific agreement. *Nature Climate Change*.

European Commission. 2021. In *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*.

D. Fallis. 2014. A functional analysis of disinformation. In *iConference Proceedings*.

Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in Russian news: a computational analysis of intricate political strategies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3570–3580, Brussels, Belgium. Association for Computational Linguistics.

C. J. Fillmore. 1976. Frame semantics and the nature of language *. *Annals of the New York Academy of Sciences*, 280.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on twitter? *SIGMETRICS Perform. Eval. Rev.*, 44(1):179–192.

Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on facebook and twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2018. Stance detection in fake news a combined feature representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Amira Ghenai and Yelena Mejova. 2018. Fake cures: User-centric modeling of health misinformation in social media. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA. Association for Computing Machinery.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. *CHI*.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Maurício Gruppi, Benjamin D. Horne, and Sibel Adalı. 2020. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna M. Howard. 2021. Mitigating racial biases in toxic language detection with an equity-based ensemble framework. *Equity and Access in Algorithms, Mechanisms, and Optimization*.

S. Hall. 1973. *Encoding and Decoding in the Television Discourse*. Media series: 1972. Centre for Cultural Studies, University of Birmingham.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. A survey on stance detection for mis- and disinformation identification. *ArXiv*, abs/2103.00242.

Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ruihong Hou, Verónica Pérez-Rosas, S. Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. *ArXiv*, abs/1909.01543.

Binxuan Huang and Kathleen M. Carley. 2020. Disinformation and misinformation on twitter during the novel coronavirus outbreak. *ArXiv*, abs/2006.04278.

F. Jahanbakhsh, Amy X. Zhang, A. Berinsky, Gordon Pennycook, David G. Rand, and D. Karger. 2021. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction*, 5:1 – 42.

Liwei Jiang, Jena D. Hwang, Chandrasekhar Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *ArXiv*, abs/2110.07574.

Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. 2021. A distributional approach to controlled text generation. *ICLR*.

Chinmay Kulkarni and Ed Chi. 2013. All the news that's fit to read: a study of social annotations for news reading. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Vivian Lai, Han Liu, and Chenhao Tan. 2020. "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Res Lett. 2017. Fake news threatens a climate literate world. *Nature Communications*, 8(15460):1.

R. Likert. 1932. A technique for the measurement of attitude scales. In *Archives of Psychology, 22 140, 55*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Yantao Jia, and Zhicheng Sheng. 2020. How does context matter? on the robustness of event detection with context-selective mask generalization. In *FINDINGS*.

I. Loshchilov and F. Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Clyde R. Miller. 1939. The techniques of propaganda. *How to Detect and Analyze Propaganda*.

Prakamya Mishra and Rohan Mittal. 2021. Neuralnere: Neural named entity relationship extraction for end-to-end climate change knowledge graph construction. *Tackling Climate Change with Machine Learning Workshop at ICML 2021*.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2021. Gpt-3 models are poor few-shot learners in the biomedical domain. *ArXiv*, abs/2109.02555.

Mohsen Mosleh, Gordon Pennycook, and David G. Rand. 2020. Self-reported willingness to share political news articles in online surveys correlates with actual sharing on twitter. *PLoS ONE*, 15.

M. Müller, M. Salathé, and P. Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *ArXiv*, abs/2005.07503.

Matthew C. Nisbet and Dietram A. Scheufele. 2009. What's next for science communication? promising directions and lingering distractions. *American Journal of Botany*, 96(10):1767–1778.

Jeppe Norregaard, Benjamin D. Horne, and Sibel Adali. 2019. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Jeff Z. Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *The Semantic Web – ISWC 2018 - 17th International Semantic Web Conference, 2018, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 669–683, Germany. Springer Verlag. 17th International Semantic Web Conference, ISWC 2018 ; Conference date: 08-10-2018 Through 12-10-2018.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Unpublished manuscript*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Zhiying (Bella) Ren, Eugen Dimant, and Maurice E. Schweitzer. 2021. Social motives for sharing conspiracy theories. *SSRN*.

Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection.

Tal Schuster, R. Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*.

Phillip R. Shaver, Judith C. Schwartz, Donald Kirson, and Cary O'Connor. 1987. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52 6:1061–86.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 3679–3686, Istanbul, Turkey. European Languages Resources Association (ELRA).

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Bertie Vidgen, Austin Botelho, David A. Broniatowski, E. Guest, M. Hall, H. Margetts, Rebekah Tromble, Zeerak Waseem, and Scott A. Hale. 2020. Detecting east asian prejudice on social media. *ArXiv*, abs/2005.03909.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *ArXiv*, abs/2002.10957.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kai-Chou Yang, Timothy Niven, and Hung-Yu Kao. 2019. Fake news detection as natural language inference. *ArXiv*, abs/1907.07347.

Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc.

Tianyi Zhang, V. Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv.*, 53(5).

A

A.1 Additional Annotation Details

We include all annotations from qualified workers in the pilots and final task as part of the dataset, discarding annotations from disqualified workers. We also removed headlines that received no annotations due to deformities in the original text (e.g. unexpected truncation) or vagueness. We paid workers at a rate of $0.4 per hit during these pilots and $.6 per hit for the second stage pilot and final task.[23] Annotators consent to doing each task by accepting it on the MTurk platform after reading a short description of what they will be asked to do.

For writer intent implications, we asked annotators if each of the 7 predefined themes was relevant to the event. If a theme was relevant, we asked annotators to provide 1-3 implications related to the chosen theme. For reader perception and action implications, we elicit 1-2 implications.

All news headlines are in English. The Poynter database contains international news originally presented in multiple languages, however news headlines contained in the database have all been translated into English.

A.1.1 Full Instructions to Annotators

> *Thanks for participating in this HIT! You will read a sentence fragment depicting an event from a news article (please note that some of these news articles may contain misinformation). The fragment may contain references to specific organizations or locations. In this case, please write an answer based on the most generic form of this reference (for example if there is a reference to the CDC, provide an answer like "government," "government organization" or "health agency," rather than writing "CDC."*
>
> *Think about what might be implied by the news event described, including the reaction it might invoke from someone reading about the news event and what the intent of the sentence fragment's author was.*
>
> *The readers' reactions and author's intent may cover multiple topic categories (for example, a sentence fragment may contain implications relating to technology and society), so when thinking about implications try to consider as many topics as possible.*

[23] We estimate this to be a fair wage of $12-$18/hr, well above minimum wage.

A.2 Post-processing of Annotations

To remove duplicate free-text annotations, we check if annotations along the same MRF dimension have a ROUGE-2 (Lin, 2004) overlap of less than .8. If two annotations have a overlap that violates this threshold, we keep one and discard the other. We also remove writer intent annotations that have a ROUGE-2 overlap of greater or equal to .8 with the headline to prevent direct copying. Due to noise in the keyword filtering approach to labeling climate-related NELA-GT headlines, we remove headlines with specific keywords referencing toxic work environments or political climates.[24]

Some "perception" annotations were more suited semantically to being "action" annotations or vice versa. If an "action" annotation is a single word categorized as a variant of a *emotion* word (Shaver et al., 1987), we reclassify it as a "perception." Conversely, if a "perception" annotation includes "want," expressing a desire for an action to happen or to do an action, we reclassify it as a "action." During this process, we also remove single word annotations that feature common misspellings.[25]

We restrict writer intent annotations to be at least three words long. Reader perception and action annotations must be at least three characters in length.

Finally, we handled missing free-text annotations. If a headline had no free-text annotations, we took this as an indicator of a low-quality example or assumed it lacked enough context for annotators to make a judgement. These invalid headlines were removed (8.5% of all headlines). If a writer intent annotation is missing, we assume the intent is ambiguous and mark it is as "unknown intent." These make up 6% of valid headlines and are not included in the overall count of implications. If a reader perception or action annotation is missing, we infer the corresponding implication from other annotated MRF variables using Table 8. Given the variables in columns 1 and 2, we randomly sample variables from 3 and 4.

[24] There may still be cornercases, but this covers the vast majority of mislabelings.

[25] While misspellings were considered during overall quality control of workers, these are difficult to handle automatically. For example, automatic spell-checkers change instances of "biden" to "widen," so we forgo automatic spellchecking.

Likelihood of Spread	Perceived Label	Potential Perceptions	Potential Actions
<3	Misinfo	'feel lied to','feel disinterested','feel disbelief','feel this is false','feel suspicious'	'fact-check this article','skip this article','check the facts','avoid sharing this article','do something else'
<3	Real or Disagree	'feel unsure','feel like they need more information to process this'	'move on to the next thing','read more', 'learn more'
>3	Any	'feel curious','feel interested','feel like this is something others might want to know about'	'talk to a friend about it','share the article','learn more','read more','try to understand'
=3	Any	'feel indifferent'	'move on to something else'

Table 8: Process for handling missing reader annotations.

Label Type	Misinfo ↓	Real ↑	Effect size
Pred	2.040	3.240	0.764
Gold	2.531	3.213	1.380

Table 9: Likelihood of news events spreading, i.e. the annotators' rating for how likely it is they would share or read the article based on the shown news event. For "Pred", we ignore headlines where annotators were unsure about the label. For this and the following tables, arrows indicate the desired direction of the score. We use Cohen's d to compute effect size.

A.3 Experimental Setup and Model Hyperparameters

All models are trained on either a single Quadro RTX 8000 or TITAN Xp GPU. Average training time for generative models ranges from approx. 1 hour per epoch for T5-base to 4 hours for GPT-2 large. Inference time for models ranges from approx. 10-20 minutes. Average training time for BERT models is approx. 30 minutes per epoch and inference time is approx. 10 minutes. We use a single final training/evaluation run and hyperparameters are manually tuned in the range of 1e-2 to 6e-6.

A.3.1 Bert Classification Models

Supervised classification models are finetuned on our corpus. Both BERT and Covid-BERT models are trained for a maximum of 30 epochs with a learning rate of 1.5e-5 and batch size of 8. Propaganda detection models are trained using the settings given in (Da San Martino et al., 2019). BERT models have 345M parameters.

A.3.2 Generative Models

For GPT-2, models are finetuned with a learning rate of 2e-5. We use a learning rate of 5e-5 for T5. For all models except GPT-2 large we use a batch size of 16. For GPT-2 large we use a batch size of 4. We use beam search with a beam size of 3 for the generation task. Generation models are trained for a maximum of 10 epochs using early stopping based on dev. loss (in the case of the GPT-2 model finetuned on cancer data we finetune for a single epoch). We optimize using AdamW (Loshchilov and Hutter, 2019) and linear warmup. Model sizes range from 124M parameters for GPT-2 small to 774M parameters for GPT-2 large.

A.4 Effect of Reader Perception on Article Sharing or Reading

Annotators tended to be cautious in reported sharing or reading behavior. We found that annotators did have a higher likelihood of sharing or reading real articles over misinformation articles (Table 9), and importantly generally claimed that they would not share or read articles that they thought were misinformation. For 1.2% of articles reported as misinformation in the training set annotators did provide a likelihood of sharing or reading $\geq$ 4. We show examples of articles that were labeled as

Type	Description	Covered by MRF
Misinformation	Misinformation is an umbrella term for news that is false or misleading.	✓
Disinformation	Unlike misinformation, disinformation assumes a malicious intent or desire to manipulate. In our framework, we focus on intent in terms of reader-perceived implications rather than questioning whether or not the writer's intentions were malicious given that it is unclear the extent to which original writers might have known article content was misleading.	Potentially
Fake News	As defined by (Allcott and Gentzkow, 2017), fake news refers to *"news articles that are intentionally and verifiably false, and could mislead readers."* (Golbeck et al., 2018) notes that fake news is a form of hoax, where the content is factually incorrect and the purpose is to mislead. This also overlaps with the definition of disinformation.	Potentially
Propaganda	Propaganda is widely held to be news that is *"an expression of opinion or action by individuals or groups, deliberately designed to influence opinions or actions of other individuals or groups with reference to predetermined ends"* (Miller, 1939). Propaganda is therefore wholly defined in terms of the intent of a writer or group of writers, and may contain factually correct content.	✓
Satire	We refer to articles written with a humorous or ironic intent as "satire." We do not explicitly cover satire in MRF, but it is possible that some misinformation articles began as satire and were misconstrued as real news.	Potentially
Real (Trusted)	We consider this to be news that is factually correct with an intent to inform. We note that while real news is distinct from most of the article types shown here, it can also function as propaganda.	✓

Table 10: Article types based on intention and perceived reliability.

Headline (Spread)	Pred/Gold
Why Companies Are Making Billions of COVID-19 Vaccine Doses That May Not Work (4.0)	Misinfo/Real
NATO's Arctic War Exercise Unites Climate Change and WWIII (4.0)	Misinfo/Real
Eat Bugs! EU Pressing member States to Promote Climate Friendly Insect Protein Diets (4.0)	Misinfo/Misinfo
Coronavirus was created in Wuhan lab and released intentionally. (5.0)	Misinfo/Misinfo

Table 11: Headlines that were labeled as misinformation by annotators and also given a high aggregated likelihood of being read or shared (spread). We show the predicted and gold labels.

"misinfo" but shared or read anyway in Table 11. While the reasoning for this is unclear, the annotators' reaction frame predictions for reader perceptions and actions may provide insight. For example, annotators were skeptical of the misinformation news event "*Coronavirus was created in Wuhan lab and released intentionally.*," but said they would share/read it anyway and provided "*readers would feel curious*" and "*readers would want to know if the wild claim has any truth to it*" as related inferences. Concerningly, this indicates even very obvious misinformation may still be shared or read by generally knowledgeable readers when it contains content they deem particularly interesting or they want to corroborate the article content with others. This aligns with a recent study of 67 million tweets (Huang and Carley, 2020) that found the "Covid as a bio-weapon started in a lab" theory is a commonly spread disinformation storyline perpetuated by bot-like accounts on Twitter.

Overall, however, we found that annotators' perception of an article as being more reliable played a positive role in their decision to share or read it.

A.5 Analysis of A/B Test

As shown by Table 12, generated writer intent implications can provide explanations that are effective at increasing reader trust in real news or decreasing trust in misinformation. However, the effect on reader trust is not always indicative of the generated intent's relevance to the headline or accuracy in capturing likely intent. Model errors like hallucinations can also decrease reader trust, as shown in the last example where the wrong state is referenced. This highlights the importance of evaluating effectiveness for both real news and misinformation.

A.6 Further Related Work

In our framework, we focus on intent in terms of implications rather than questioning whether or not the writer's intentions were malicious given that it is unclear the extent to which original writers might have known article content was misleading. We summarize common definitions for news reliability in Table 10).

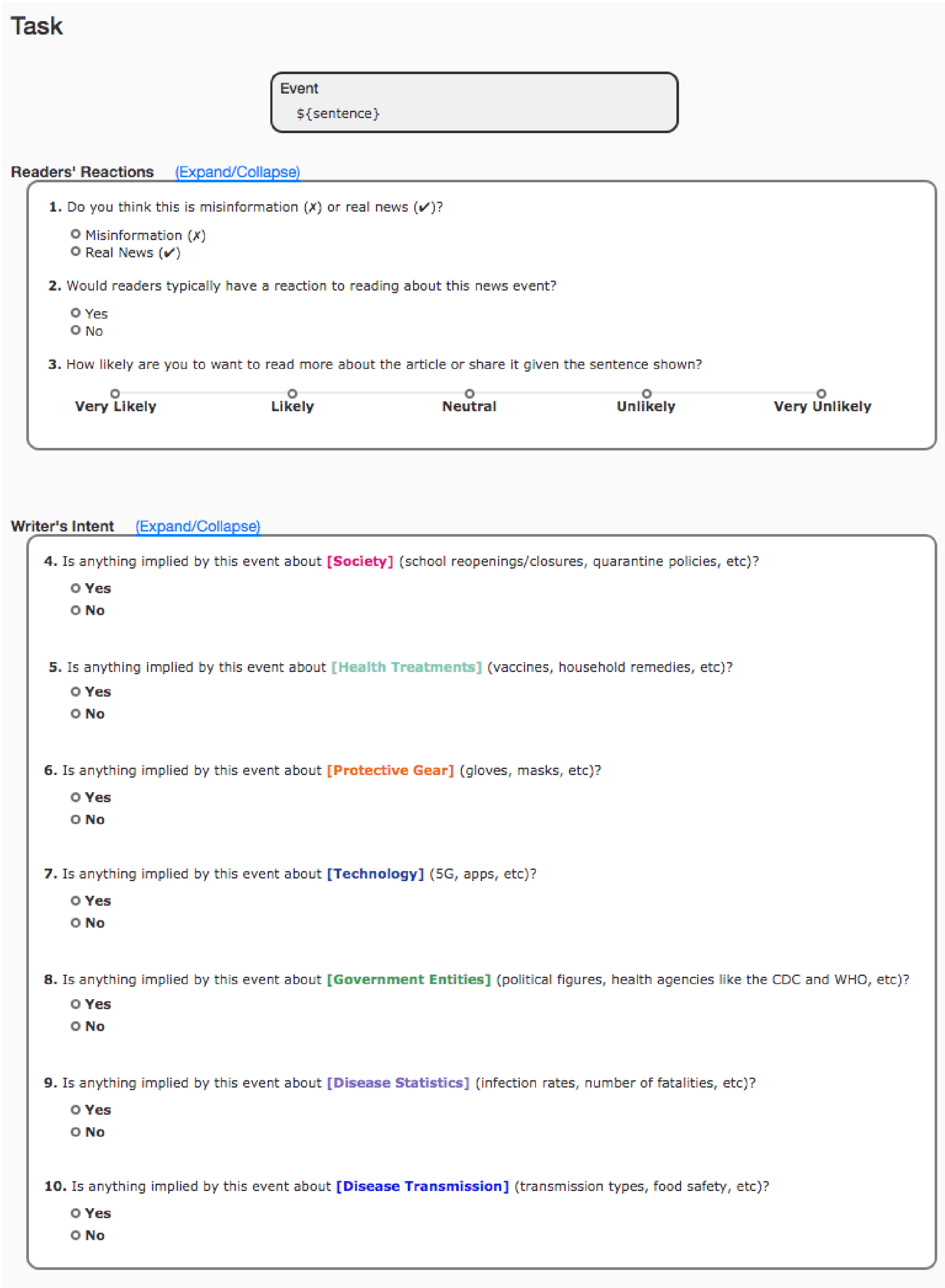
Rhetorical Framing of News. Prior work on rhetorical framing (e.g. Nisbet and Scheufele, 2009; Card et al., 2015; Field et al., 2018) has noted the significant role *media frames* play in shaping public perception of social and political issues, as well as the potential for misleading representations of events in news media. However, past formalisms for rhetorical framing that rely on common writing or propaganda techniques (e.g. *appeal to fear* or *loaded language*, Da San Martino et al., 2019) do not explicitly model impact. To that end, we propose a formalism focusing on readers' perception of the writers' intention, rather than specific well-known techniques.

Misinformation Detection. There has been work on integration of knowledge graphs (Pan et al., 2018) and framing detection as a NLI task (Yang et al., 2019). Zellers et al. (2019) show the effectiveness of using large-scale neural language modeling to detect machine-generated misinformation. Recent work has also highlighted the importance of understanding the impact from misinformation, particularly in health domains (Dharawat et al., 2020; Ghenai and Mejova, 2018). Zhou and Zafarani (2020) and Hardalov et al. (2021) provide comprehensive surveys of misinformation detection methods. Our work is related to stance detection (Ghanem et al., 2018), however our pragmatic frames go beyond understanding the stance of a reader and explicitly capture how reader perceptions affect their actions.

Countering Misinformation. It has been noted in prior work that sharing behavior reported in MTurk crowdsourced studies matches behavior in-the-wild (Mosleh et al., 2020). (Yaqub et al., 2020; Lai et al., 2020) show the effectiveness of credibility indicators to persuade readers to decrease their trust in false information. (Jahanbakhsh et al., 2021) show that having users assess accuracy of news at sharing time and providing rationales for their decisions decreases likelihood of false information being shared.

Headline	Generated Writer Intent (Model)	Shift in Trust	Gold Label
Every day in Germany more people die because of wrong medical treatment, misuse of drugs or hospital germs than of Covid-19	The writer is implying that the pandemic isn't that bad (T5-large)	Decreases Trust	Misinfo
NYC COVID-19 Deaths During Peak Rivaled 1918 Flu Fatalities	The writer is implying that the pandemic is dangerous (T5-large)	Increases Trust	Real
PCR Tests cannot show the novel coronavirus.	The writer is implying that covid testing is unreliable (GPT-2 large)	Decreases Trust	Misinfo
Alaska's new climate threat: tsunamis linked to melting permafrost	The writer is implying that climate change is real (GPT-2 large)	Increases Trust	Real
"Nearly half of (Missouri) counties have not reported positive (COVID-19) cases."	The writer is implying that covid is not spreading in Missouri (GPT-2 large)	Decreases Trust	Misinfo
Can the catastrophic fires bring some sanity to Australian climate politics?	The writer is implying that wildfires in australia are a result of climate change (GPT-2 large)	Increases Trust	Real
Wisconsin is "clearly seeing a decline in COVID infections".	The writer is implying that covid is not spreading in florida (GPT-2 large)	Decreases Trust	Misinfo

Table 12: Examples where generated writer intent implications are effective at changing perceived trustworthiness of news headlines.



Task

Event
${sentence}

Readers' Reactions (Expand/Collapse)

1. Do you think this is misinformation (✗) or real news (✔)?

- Misinformation (✗)
- Real News (✔)

2. Would readers typically have a reaction to reading about this news event?

- Yes
- No

3. How likely are you to want to read more about the article or share it given the sentence shown?

Very Likely · **Likely** · **Neutral** · **Unlikely** · **Very Unlikely**

Writer's Intent (Expand/Collapse)

4. Is anything implied by this event about **[Society]** (school reopenings/closures, quarantine policies, etc)?

- **Yes**
- **No**

5. Is anything implied by this event about **[Health Treatments]** (vaccines, household remedies, etc)?

- **Yes**
- **No**

6. Is anything implied by this event about **[Protective Gear]** (gloves, masks, etc)?

- **Yes**
- **No**

7. Is anything implied by this event about **[Technology]** (5G, apps, etc)?

- **Yes**
- **No**

8. Is anything implied by this event about **[Government Entities]** (political figures, health agencies like the CDC and WHO, etc)?

- **Yes**
- **No**

9. Is anything implied by this event about **[Disease Statistics]** (infection rates, number of fatalities, etc)?

- **Yes**
- **No**

10. Is anything implied by this event about **[Disease Transmission]** (transmission types, food safety, etc)?

- **Yes**
- **No**

Figure 2: Layout of annotation task for collecting Covid-related MRF data.