Structurizing Misinformation Stories via Rationalizing Fact-Checks

Shan Jiang and **Christo Wilson**
Northeastern University, USA
{sjiang, cbw}@ccs.neu.edu

Abstract

Misinformation has recently become a well-documented matter of public concern. Existing studies on this topic have hitherto adopted a coarse concept of misinformation, which incorporates a broad spectrum of story types ranging from political conspiracies to misinterpreted pranks. This paper aims to structurize these misinformation stories by leveraging fact-check articles. Our intuition is that key phrases in a fact-check article that identify the misinformation type(s) (*e.g.,* doctored images, urban legends) also act as rationales that determine the verdict of the fact-check (*e.g.,* false). We experiment on rationalized models with domain knowledge as weak supervision to extract these phrases as rationales, and then cluster semantically similar rationales to summarize prevalent misinformation types. Using archived fact-checks from Snopes.com, we identify ten types of misinformation stories. We discuss how these types have evolved over the last ten years and compare their prevalence between the 2016/2020 US presidential elections and the H1N1/COVID-19 pandemics.

1 Introduction

Misinformation has raised increasing public concerns globally, well-documented in Africa (Ahinkorah et al., 2020), Asia (Kaur et al., 2018), and Europe (Fletcher et al., 2018). In the US, "fake news" accounted for 6% of all news consumption during the 2016 US presidential election (Grinberg et al., 2019). Years later, 29% of US adults in a survey believed that the "exaggerated threat" of the COVID-19 pandemic purposefully damaged former US president Donald Trump (Uscinski et al., 2020), and 77% of Trump's supporters believed "voter fraud" manipulated the 2020 US presidential election in spite of a complete lack of evidence (Pennycook and Rand, 2021).

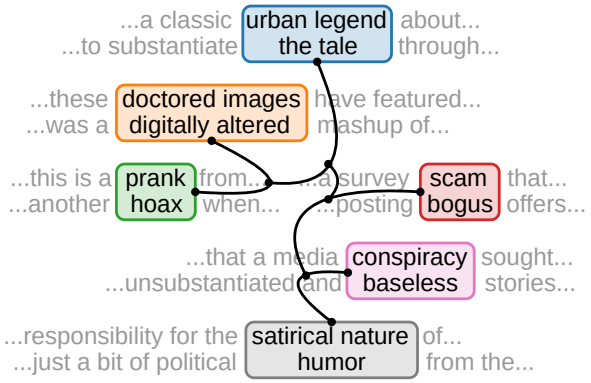

Figure 1: **A snippet of the misinformation structure.** Each line is a snippet from a fact-check. Key phrases identifying the misinformation types are highlighted. Phrases with similar semantics are clustered in colored boxes. This structure is a sample of our final results.

As such misinformation continues to threaten society, researchers have started investigating this multifaceted problem, from understanding the socio-psychological foundations of susceptibility (Bakir and McStay, 2018) and measuring public responses (Jiang and Wilson, 2018; Jiang et al., 2020b), to designing detection algorithms (Shu et al., 2017) and auditing countermeasures for online platforms (Jiang et al., 2019, 2020c).

These studies mostly adopted the term "misinformation" as a coarse concept for any false or inaccurate information, which incorporates a broad spectrum of misinformation stories, *e.g.,* political conspiracies to misinterpreted pranks. Although misinformation *types* have been theorized and categorized by practitioners (Wardle, 2017), there is, to our knowledge, no empirical research that has systematically measured these prevalent types of misinformation stories.

This paper aims to unpack the coarse concept of misinformation and structurize it to fine-grained story types (as illustrated in Figure 1). We conduct

this query through an empirical lens and ask the question: *what are the prevalent types of misinformation stories in the US over the last ten years?*

The answer to our question is buried in archived fact-checks, which are specialized news articles that verify factual information and debunk false claims by presenting contradictory evidence (Jiang et al., 2020a). As a critical component of their semi-structured journalistic style, fact-checks often embed the (mis)information type(s) within their steps of reasoning. For example, consider the following snippet from a Snopes.com fact-check with a verdict of **false** (Evon, 2019):

> "...For instance, some started sharing a **doctored photograph** of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore)..."

The key phrase **doctored photograph** in the snippet identifies the misinformation type of the fact-checked story. Additional example phrases are highlighted in Figure 1. With a large corpus of fact-checks, these phrases would accumulate and reveal prevalent types of misinformation stories.

Extracting these phrases is a computational task. Our intuition is that such phrases in a fact-check also act as *rationales* that determine the verdict of the fact-check. In the previous example, the verdict is **false** in part *because* the story contains a **doctored photograph**. Therefore, a neural model that predicts the verdict of a fact-check would also use the misinformation types as rationales.

To realize this intuition, we experiment on existing rationalized neural models to extract these phrases (Lei et al., 2016; Jain et al., 2020), and, to target specific kinds of rationales, we additionally propose to include domain knowledge as weak supervision in the rationalizing process. Using public datasets as validation (Zaidan et al., 2007; Carton et al., 2018), we evaluate the performance variation of different rationalized models, and show that including domain knowledge consistently improves the quality of extracted rationales.

After selecting the most appropriate method, we conduct an empirical investigation of prevalent misinformation types. Using archived fact-checks from Snopes.com, spanning from its founding in 1994 to 2021, we extract rationales by applying the selected model with theorized misinformation types for weak supervision (Wardle, 2017), and then cluster rationales based on their semantic similarity to summarize prevalent misinformation types. We identify ten types of misinformation stories, a preview of which are shown in Figure 1.

Using our derived lexicon of these clustered misinformation stories, we then explore the evolution of misinformation types over the last ten years. Our key findings include: increased prevalence of conspiracy theories, fabricated content, and digital manipulation; and decreased prevalence of legends and tales, pranks and jokes, mistakes and errors, etc. We also conducted two case studies on notable events that involve grave misinformation. From the case study of US presidential elections, we observe that the most prevalent misinformation type for both the 2016 and 2020 elections is fabricated content, while the 2016 election has more hoaxes and satires. From the case study of pandemics, our results show that the H1N1 pandemic in 2009 has more legends and tales, while the COVID-19 pandemic attracts more conspiracy theories.

The code and data used in the paper are available at: `https://factcheck.shanjiang.me`.

2 Related Work

There is a rich literature that has studied the online misinformation ecosystem from multiple perspectives (Del Vicario et al., 2016; Lazer et al., 2018). Within the computational linguistics community, from an audiences' perspective, Jiang and Wilson (2018) found that social media users expressed different linguistic signals when responding to false claims, and the authors later used these signals to model and measure (dis)beliefs in (mis)information (Jiang et al., 2020b; Metzger et al., 2021). From a platforms' perspective, researchers have assisted platforms in designing novel misinformation detection methods (Wu et al., 2019; Lu and Li, 2020; Vo and Lee, 2018, 2020), as well as audited existing misinformation intervention practices (Robertson et al., 2018; Jiang et al., 2019, 2020c; Hussein et al., 2020).

In this work, we study another key player in the misinformation ecosystem, *storytellers*, and investigate the prevalent types of misinformation told to date. From the storytellers' perspective, Wardle (2017) theorized several potential misinformation types (*e.g.,* satire or parody, misleading content, and false connection), yet no empirical evidence has been connected to this typology. Additionally, researchers have investigated specific types of mis-

information as case studies, *e.g.,* state-sponsored disinformation (Starbird et al., 2019; Wilson and Starbird, 2020), fauxtography (Zannettou et al., 2018; Wang et al., 2021), and conspiracy theories (Samory and Mitra, 2018; Phadke et al., 2021). In this paper, we aim to structurize these misinformation stories to theorized or novel types.

3 Rationalized Neural Models

Realizing our intuition (as described in § 1) requires neural models to (at least shallowly) reason about predictions. In this section, we introduce existing rationalized neural models and propose to include domain knowledge as weak supervision in the rationalizing process. We then experiment with public datasets and lexicons for evaluation.

3.1 Problem Formulation

In a standard text classification problem, each instance is in a form of $(\boldsymbol{x}, \boldsymbol{y})$. $\boldsymbol{x} = [\boldsymbol{x}^i] \in V_x^l$ is the input token sequence of length l, where V_x is the vocabulary of the input and i is the index of each token $\boldsymbol{x}^i$. $\boldsymbol{y} \in \{0,1\}^m$ is the binary label of length m. Rationalization requires a model to output the prediction $\hat{\boldsymbol{y}}$ together with a binary mask $\boldsymbol{z} = [\boldsymbol{z}^i] \in \{0,1\}^l$ of input length l, indicating which tokens are used (*i.e.,* $\boldsymbol{z}^i = 1$) to make the decision. These tokens are called *rationales*.

Hard rationalization requires a model to directly output $\boldsymbol{z}$. Initially proposed by Lei et al. (2016), the model first passes the input $\boldsymbol{x}$ to a tagger[1] module and samples a binary mask $\boldsymbol{z}$ from a Bernoulli distribution, *i.e.,* $\boldsymbol{z} \sim \text{Tagger}(\boldsymbol{x})$, and then uses only unmasked tokens to make a prediction of $\boldsymbol{y}$, *i.e.,* $\hat{\boldsymbol{y}} = \text{Predictor}(\boldsymbol{z}, \boldsymbol{x})$.[2]

The loss function of this method contains two parts. The first part is a standard loss for the prediction $L_y(\hat{\boldsymbol{y}}, \boldsymbol{y})$, which can be realized using common classification loss, *e.g.,* cross entropy. The second part is a loss $L_z(\boldsymbol{z})$[3] aiming to regularize $\boldsymbol{z}$ and encourage conciseness and contiguity of rationale selection, formulated by Lei et al. (2016). Recent work proposed to improve the initial model with an adversarial component (Yu et al., 2019; Carton et al., 2018). Combining these parts together, the model is trained end-to-end using reinforce-style estimation (Williams, 1992), as sampling rationales is a non-differentiable computation. The modules of hard rationalization are illustrated in Figure 2.

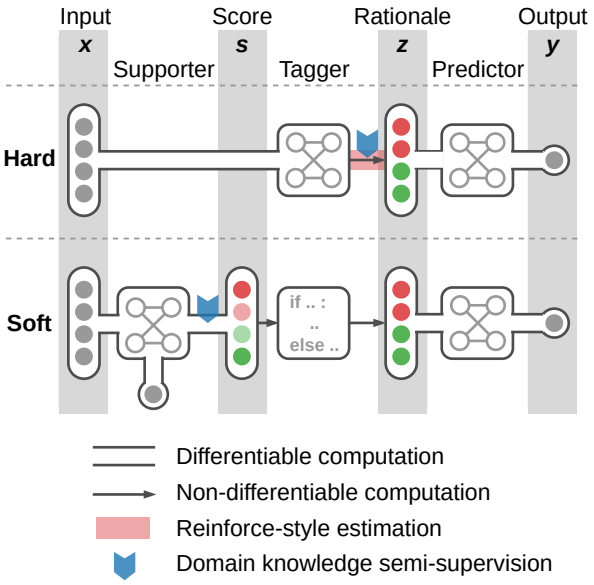

Figure 2: **Hard and soft rationalization methods.** Hard rationalization is an end-to-end model that first uses input $\boldsymbol{x}$ to generate rationales $\boldsymbol{z}$, and then uses unmasked tokens to predict $\boldsymbol{y}$. Soft rationalization is a three-phased model that first uses input $\boldsymbol{x}$ to predict $\boldsymbol{y}$ and outputs importance scores $\boldsymbol{s}$, then binarizes $\boldsymbol{s}$ to rationales $\boldsymbol{z}$, and finally uses unmasked tokens to predict $\boldsymbol{y}$ again as evaluation for faithfulness.

Soft rationalization, in contrast, allows a model to first output a continuous version of importance scores $\boldsymbol{s} = [\boldsymbol{s}^i] \in \mathbb{R}^l$, and then binarize it to get $\boldsymbol{z}$. Initially formalized by Jain et al. (2020) as a multiphase method, the model first conducts a standard text classification using a supporter module $\hat{\boldsymbol{y}} = \text{Supporter}(\boldsymbol{x})$ and outputs importance scores $\boldsymbol{s}$, then binarizes $\boldsymbol{s}$ using a tagger module, *i.e.,* $\boldsymbol{z} = \text{Tagger}(\boldsymbol{s})$, and finally uses only unmasked tokens of $\boldsymbol{x}$ to make another prediction $\hat{\boldsymbol{y}}$ to evaluate the faithfulness of selected rationales.[4]

These three modules are trained separately in three phases.[5] Since the supporter and predictor are standard text classification modules the only loss needed is for the prediction $L_y(\hat{\boldsymbol{y}}, \boldsymbol{y})$. This method is more straightforward than the hard rationalization method, as it avoids non-differentiable com-

[1] This module was named *generator* by Lei et al. (2016). We name it *tagger* to distinguish it from the NLG problem.

[2] This module was named *encoder* by Lei et al. (2016). We name it *predictor*, consistent with Yu et al. (2019), to distinguish it from the encoder-decoder framework.

[3] $L_z(\boldsymbol{z})$ is a simplified term; we discuss its detailed implementation in Appendix § A.

[4] The second and third modules were named *extractor* and *classifier* by Jain et al. (2020). We continue using *tagger* and *predictor* to align with the hard rationalization method.

[5] Tagger is often flexibly designed as a rule-based algorithm, therefore no training is needed.

putations and the instability induced by reinforce-style estimation. The modules of soft rationalization are also illustrated in Figure 2.

The popular *attention* mechanism (Bahdanau et al., 2014) provides built-in access to $\boldsymbol{s}$. Although there have been debates on the properties achieved by attention-based explanations (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019), rationales extracted by straightforward rules on attention weights were demonstrated as comparable to human-generated rationales (Jain et al., 2020). Additionally, in our use case we only need the rationales themselves as key phrases and do not require them to faithfully predict $\boldsymbol{y}$, therefore the last predictor module can be omitted.

3.2 Domain Knowledge as Weak Supervision

Both hard and soft rationalization methods can be trained with or without supervision *w.r.t.* rationales $\boldsymbol{z}$ (DeYoung et al., 2020)[6]. When rationales are selected in an unsupervised manner, the model would intuitively favor rationales that are most informative to predict the corresponding label as a result of optimizing the loss function. This could result in some undesirable rationales in our case: for example, certain entities like "COVID-19" or "Trump" that are highly correlated with misinformation would be selected as rationales even though they do not suggest any misinformation types. Therefore, we propose to weakly supervise[7] the rationalizing process with domain knowledge to obtain specific, desired types of rationales.

Assuming a lexicon of vocabulary V_d as domain knowledge, we reprocess the input and generate weak labels for rationales $\boldsymbol{z_d} = [\boldsymbol{z_d^i}] \in \{0,1\}^l$ where $\boldsymbol{z_d^i} = 1$ (*i.e.,* unmasked) if $\boldsymbol{x}^i \in V_d$ and $\boldsymbol{z_d^i} = 0$ (*i.e.,* masked) otherwise. Then, we include an additional loss item $L_d(\boldsymbol{z}, \boldsymbol{z_d})$ or $L_d(\boldsymbol{s}, \boldsymbol{z_d})$ for the hard or soft rationalization method.

Combining the loss items together, the objective for the end-to-end hard rationalization model is:

$$\min_{\theta} L_y(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \lambda_z L_z(\boldsymbol{z}) + \lambda_d L_d(\boldsymbol{z}, \boldsymbol{z_d}),$$

where θ contains the parameters to estimate and $\lambda_{(\cdot)}$ are hyperparameters weighting loss items.

Similarly, the objective function for the first phase of soft rationalization is:

$$\min_{\theta} L_y(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \lambda_d L_d(\boldsymbol{s}, \boldsymbol{z_d}).$$

[6]They are trained with supervision *w.r.t.* the label $\boldsymbol{y}$.

[7]Since there is inherently no ground-truth of misinformation types in fact-check articles.

3.3 Experiments on Public Datasets

We conduct experiments on public datasets to evaluate the performance of hard and soft rationalization methods, particularly for our needs, and confirm that including domain knowledge as weak supervision helps with the rationalizing process.

Datasets selection. An ideal dataset for our models should meet the following requirements: **(a)** formulated as a text classification problem, **(b)** annotated with human rationales, and **(c)** can be associated with high quality lexicons to obtain domain knowledge. We select two datasets based on these criteria: the **movie reviews** dataset released by Pang et al. (2002) and later annotated with rationales by Zaidan et al. (2007), which contains 2K movie reviews labeled with positive or negative sentiments; and the **personal attacks** dataset released by Wulczyn et al. (2017) and later annotated with rationales by Carton et al. (2018), which contains more than 100K Wikipedia comments labeled as personal attacks or not.

Domain knowledge. For the sentiment analysis on movie reviews, we use the EmoLex lexicon released by Mohammad and Turney (2013), which contains vocabularies of positive and negative sentiments. For identifying personal attacks, we use a lexicon released by Wiegand et al. (2018), which contains a vocabulary of abusive words. With corresponding vocabularies, we generate weak rationale labels $\boldsymbol{z_d}$ for each dataset.

Evaluation metrics. We choose binary precision $\Pr(\boldsymbol{z})$ to evaluate the quality of extracted rationales, because **(a)** a perfect recall can be trivially achieved by selecting *all* tokens as rationales,[8] and **(b)** our case of identifying key phrases requires concise rationales. Additionally, we measure the average percentage of selected rationales over the input length $\%(\boldsymbol{z})$. For predictions, we use macro $\mathrm{F}_1(\boldsymbol{y})$ as the evaluation metric as well as the percentage of information used $\%(\boldsymbol{x})$ to make the prediction.

Experimental setup and results. The train, dev, and test sets are pre-specified in public datasets. We optimize hyperparameters for $\mathrm{F}_1(\boldsymbol{y})$ on the dev sets, and only evaluate rationale quality $\Pr(\boldsymbol{z})$ *after* a model is decided. We discuss additional implementation details (*e.g.,* hyperparameters, loss functions, module cells) in Appendix § A.

[8]We later show that this is the default model behavior if rationale selection is under-regularized.

		Movie reviews (Zaidan et al., 2007)				Personal attacks (Carton et al., 2018)			
		Pr($\boldsymbol{z}$)	%($\boldsymbol{z}$)	$F_1(\boldsymbol{y})$	%($\boldsymbol{x}$)	Pr($\boldsymbol{z}$)	%($\boldsymbol{z}$)	$F_1(\boldsymbol{y})$	%($\boldsymbol{x}$)
h_0	**Hard rationalization**	0.37	2.7%	0.72	2.7%	0.17	32.5%	0.73	32.5%
h_1	w/ Domain knowledge	0.38	3.7%	0.72	3.7%	0.22	16.9%	0.73	16.9%
h_2	w/o Rationale regularization	0.31	99.9%	0.92	99.9%	0.19	99.9%	0.82	99.9%
h_3	w/ Adversarial components	0.33	2.5%	0.70	2.5%	0.22	14.9%	0.75	14.9%
s_0	**Soft rationalization**	0.58	3.7%	0.91	100%	0.35	16.9%	**0.82**	100%
s_1	w/ Domain knowledge	**0.62**	3.7%	**0.92**	100%	**0.39**	16.9%	**0.82**	100%
s_2	w/ Half rationales	0.64	1.9%	0.92	100%	0.46	8.4%	0.82	100%
s_3	w/ Double rationales	0.55	7.4%	0.92	100%	0.31	33.8%	0.82	100%

Table 1: **Evaluation results for hard and soft rationalization methods.** Our experiments show that: **(a)** hard rationalization requires a sensitive hyperparameter λ_z to regularize rationales (h_2 to h_0); **(b)** soft rationalization achieves the best $F_1(\boldsymbol{y})$ overall, but Pr($\boldsymbol{z}$) depends on the rationale extraction approach (s_2/s_3 to s_0); **(c)** domain knowledge as weak supervision improves Pr($\boldsymbol{z}$) for both hard (h_1 to h_0) and soft (s_1 to s_0) rationalization while maintaining similar %($\boldsymbol{z}$) and $F_1(\boldsymbol{y})$; **(d)** soft rationalization achieves better Pr($\boldsymbol{z}$) in a fair comparison (s_1 to h_1).

The evaluation results for all our experiments on test sets are reported in Table 1, indexed with h_0-h_3 and s_0-s_3. We report the evaluation results on dev sets in Appendix § B.

Regularization for hard rationalization. h_0 and h_2 are our re-implementation of Lei et al. (2016), varying the rationale regularization hyperparameter λ_z. Our experiments show that λ_z is a crucial choice. When a small λ_z is chosen (*i.e.,* rationales are under-regularized), the model has a tendency to utilize all the available information to optimize the predictive accuracy. In h_2, we set $\lambda_z = 0$ and the model selects 99.9% of tokens as rationales while achieving the best $F_1(\boldsymbol{y})$ overall, which is an undesirable outcome in our case. Therefore, we increases λ_z so that only small parts of tokens are selected as rationales in h_0. However, echoing Jain et al. (2020), the output when varying λ_z is sensitive and unpredictable, and searching for this hyperparameter is both time-consuming and energy-inefficient. We also run an experiment h_3 with the additional adversarial component proposed in (Carton et al., 2018; Yu et al., 2019), and the evaluation metrics are not consistently improved compared to h_0.

Binarization for soft rationalization. s_0, s_2 and s_3 are our re-implementation of Jain et al. (2020). For soft rationalization, rationales are selected (*i.e.,* binarized) after the supporter module is trained in phase one, therefore s_0-s_3 utilize 100% of the tokens by default, and achieve the best $F_1(\boldsymbol{y})$ overall. We implement a straightforward approach to select rationales by setting a threshold t and make $\boldsymbol{z}^i = 1$ (*i.e.,* unmasked) if the importance score $\boldsymbol{s}^i > t$ and $\boldsymbol{z}^i = 0$ (*i.e.,* masked) otherwise. Intuitively, increasing t corresponds to less selected rationales, and therefore increasing Pr($\boldsymbol{z}$). To confirm, in s_2, we increase t until %($\boldsymbol{z}$) is exactly half of s_0. Similarly, decreasing t corresponds to more selected rationales, and therefore decreasing Pr($\boldsymbol{z}$). In s_3, we decrease t until %($\boldsymbol{z}$) is exactly double of s_0.

Is domain knowledge helpful? h_1 and s_1 include domain knowledge as weak supervision. Our results show that domain knowledge improves Pr($\boldsymbol{z}$) for both hard (h_1 to h_0) and soft (s_1 to s_0) rationalization methods and on both dataset, while maintaining similar %($\boldsymbol{z}$) and $F_1(\boldsymbol{y})$. The improvements are more substantial for soft rationalization.

Hard vs. soft rationalization. To fairly compare hard and soft rationalization methods, we choose the threshold t to keep %($\boldsymbol{z}$) the same for h_1 and s_1.[9] Our experiments show that soft rationalization weakly supervised by domain knowledge achieves better Pr($\boldsymbol{z}$) on both datasets, and therefore we chose it for rationalizing fact-checks.

4 Rationalizing Fact-Checks

After determining that soft rationalization is the most appropriate method, we apply it to extract rationales from fact-checks. In this section, we introduce the dataset we collected from Snopes.com and conduct experiment with fact-checks to structurize misinformation stories.

4.1 Data Collection

Snopes.com is a renowned fact-checking website, certified by the International Fact-Checking Network as non-partisan and transparent (Poynter,

[9] We can easily and accurately manipulate %($\boldsymbol{z}$) for soft rationalization by adjusting t; conversely, the impact of adjusting λ_z in hard rationalization is unpredictable.

2018). We collect HTML webpages of fact-check articles from Snopes.com, spanning from its founding in 1994 to the beginning of 2021.

Preprocess and statistics. We first preprocess collected fact-checks by extracting the main article content and verdicts from HTML webpages using a customized parser, and tokenizing the content with NLTK (Bird, 2006). The preprocessing script is included in our released codebase.

After preprocessing, the median sequence length of fact-checks is 386 tokens, and 88.6% of fact-checks containing $\leq$1,024 tokens. Jiang et al. (2020a) found that the most informative content in fact-checks tended to be located at the head or the tail of the article content. Therefore, we set the maximum sequence length to 1,024 and truncate over-length fact-checks.

Next, we label each fact-check with a binary label depending on its verdict: (truthful) information if the verdict is at least **mostly true** and misinformation otherwise, which results in 2,513 information and 11,183 misinformation instances.

Additionally, we preemptively mask tokens that are the exact words as its verdict (*e.g.,* "rate it as false" to "rate it as [MASK]"),[10] otherwise predicting the verdict would be trivial and the model would copy overlapping tokens as rationales.

Domain knowledge for misinformation types. The domain knowledge comes from two sources: **(a)** the misinformation types theorized by Wardle (2017), *e.g.,* misleading or fabricated content; and **(b)** certain variants of verdicts from Snopes.com such as satire or scam (Snopes.com, 2021a). We combine these into a small vocabulary V_d containing 12 words, listed in Appendix § A.

4.2 Experiments and Results

We randomly split the fact-checks to 80% train, 10% dev, and 10% test sets, and adjust hyperparameters to optimize $F_1(\boldsymbol{y})$ on dev set. For initialization, we train word embeddings using Gensim (Rehurek and Sojka, 2011) on the entire corpus. The final model achieves $F_1(\boldsymbol{y}) = 0.75/0.74$ on the test set with/without domain knowledge.

Clustering rationales. To systematically understand extracted rationales, we cluster these rationales based on semantic similarity. For each rationale, we average word embeddings to represent the embedding of the rationale, and then run a hierarchical clustering for these embeddings. The hierarchical clustering uses cosine similarity as the distance metric, commonly used for word embeddings (Mikolov et al., 2013), and the complete link

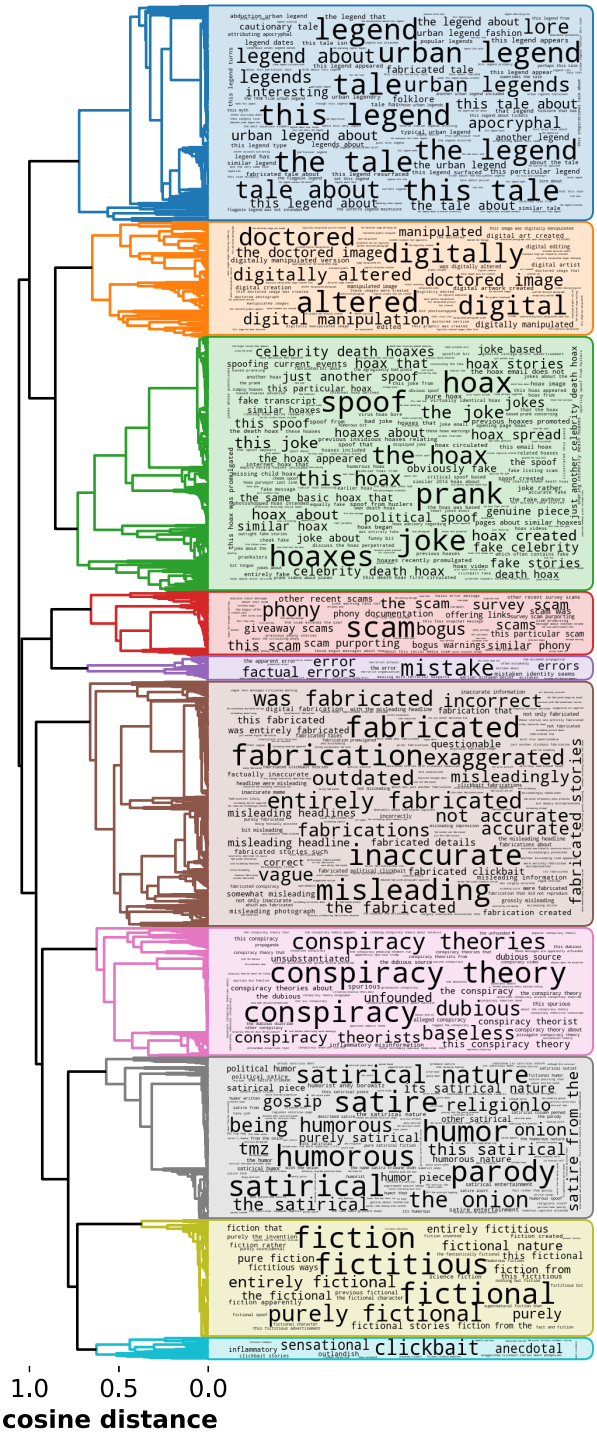

Figure 3: **Structure of misinformation types.** The ten identified clusters (colored) offer empirical confirmation of theorized misinformation types, contain novel fine-grained clusters, and reorganize the structure of misinformation stories.

[10] Verdicts from Snopes.com are structured HTML fields that can be easily parsed.

method (Voorhees, 1986) to obtain a relatively balanced linkage tree.

The results from the clustering are shown in Figure 3. From the root of the dendrogram, we can traverse its branches to find clusters until we reach a sensible threshold of cosine distance, and categorize the remaining branches and leaf nodes (*i.e.,* rationales) to multiple clusters. Figure 3 shows an example visualization that contains ten clusters of rationales that are semantically similar to the domain knowledge, and leaf nodes in each cluster are aggregated to plot a word cloud, with the frequency of a node encoded as the font size of the phrase.

Note that rationales extracted from soft rationalization are dependent on the chosen threshold t to binarize importance scores. The example in Figure 3 uses a threshold of $t = 0.01$. Varying the threshold would affect extracted rationales but mostly the ones with low prevalence, and these rare rationales also correspond to small font sizes in the word cloud. Therefore, the effect from varying t would be visually negligible in Figure 3.

Structure of misinformation stories. We make the following observations from the ten clusters of misinformation types identified in Figure 3.

First, the clusters empirically *confirm* existing domain knowledge in V_d. Certain theorized misinformation types, such as satires and parodies ■ from (Wardle, 2017), are identified as individual clusters from fact-checks.

Second, the clusters *complement* V_d with additional phrases describing (semantically) similar misinformation types. For example, our results add "humor" and "gossip" to the same category as satires and parodies ■ and add "tales" and "lore" to the same category as legends ■. This helps us grasp the similarity between misinformation types, and also enriches the lexicon V_d, which proves useful for subsequent analysis in § 5.

Third, we *discover* novel, fine-grained clusters that are not highlighted in V_d. There are multiple possible explanations as to why these misinformation types form their own clusters. Conspiracy theories ■ are often associated with intentional political campaigns (Samory and Mitra, 2018) which can affect their semantics when referenced in fact-checks. In contrast, digital alteration ■ is a relatively recent misinformation tactic that has been enabled by technological developments such as FaceSwap (Korshunova et al., 2017) and DeepFake (Westerlund, 2019). Hoaxes and pranks ■ often have a mischievous intent that distinguishes them from other clusters. Other new clusters include clickbait with inflammatory and sensational language ■ and entirely fictional content ■.

Fourth, the clusters *reorganize* the structure of these misinformation types based on their semantics, *e.g.,* fabricated and misleading content ■ belongs to two types of misinformation in (Wardle, 2017), while in our results they are clustered together. This suggests that the semantic distance between fabricated and misleading content is less than the chosen similarity threshold, at least when these misinformation types are referred to by fact-checkers when writing articles.

Finally, the remaining words in V_d are also found in our rationales. However, due to low prevalence, they are not visible in Figure 3 and do not form their own clusters.

5 Evolution of Misinformation

In this section, we leverage the clusters of misinformation types identified by our method as a lexicon and apply it back to the our original fact-check dataset. Specifically, we analyze the evolution of misinformation types over the last ten years and compare misinformation trends around major real-world events.

Evolution over the last ten years. We first explore the evolution of misinformation over time. We map each fact-check article with one or more corresponding misinformation types identified by our method, and then aggregate fact-checks by year from before 2010[11] to the end of 2020 to estimate the relative ratio of each misinformation type.

As shown in Figure 4,[12] the prevalence of certain misinformation types on Snopes.com has drastically changed over the last ten years.

Heavily politicized misinformation types, such as digitally altered or doctored images or photographs ■, fabricated and misleading content ■, and conspiracy theories ■ have nearly doubled in relative ratios over the last ten years. In contrast, the prevalence of (arguably) less politicized stories, such as legends and tales ■, hoaxes and pranks ■, and mistakes and errors ■ have decreased.

These trends may be a proxy for the underlying prevalence of different misinformation types within the US. Studies that measure political ideologies

[11] Since there are relatively few fact-checks before 2010, we aggregate them together to the year 2010.

[12] 95% confidence intervals.

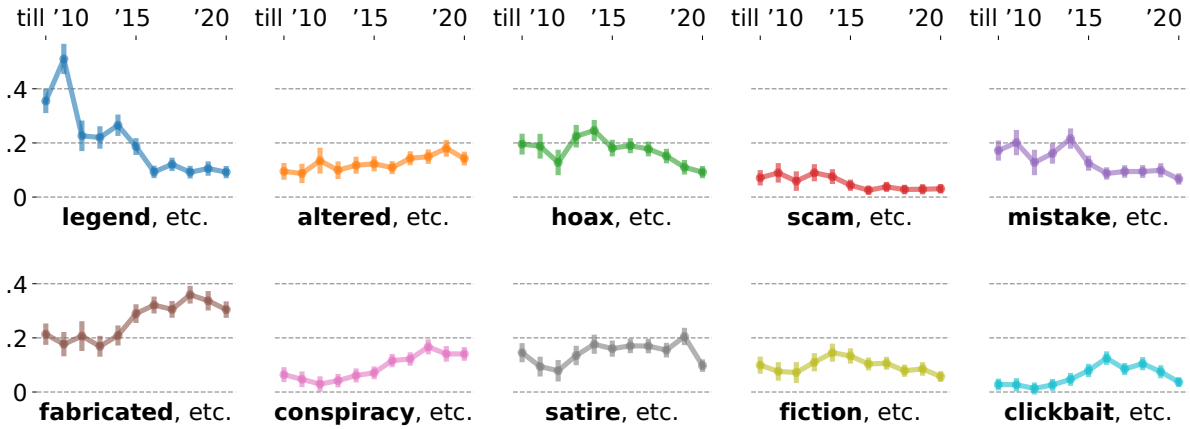

Figure 4: **Evolution of misinformation over the last ten years.** Conspiracy theories, fabricated content, and digital manipulation have increased in prevalence. The prevalence of (arguably) less politicized stories (*e.g.,* legends and tales, pranks and jokes, mistakes and errors) has decreased. (95% confidence intervals.)

expressed online have documented increasing polarization over time (Chinn et al., 2020; Baumann et al., 2020), which could explain increased ratios of such heavily politicized misinformation. Additionally, the convenience offered by modern digital alteration software and applications (Korshunova et al., 2017; Westerlund, 2019) provides a gateway to proliferating manipulated images or photographs in the misinformation ecosystem.

Alternatively, these trends may reflect shifts in Snopes.com's priorities. The website, launched in 1994, was initially named *Urban Legends Reference Pages*. Since then it has grown to encompass a broad spectrum of subjects. Due to its limited resources, fact-checkers from Snopes.com only cover a subset of online misinformation, and their priority is to "fact-check whatever items the greatest number of readers are asking about or searching for at any given time (Snopes.com, 2021b)."[13] Given the rising impact of political misinformation in recent years (Zannettou et al., 2019, 2020), such misinformation could reach an increasing number of Snopes.com readers, and therefore the website may dedicate more resources to fact-checking related types of misinformation. Additionally, Snopes.com has established collaborations with social media platforms, *e.g.,* Facebook (Green and Mikkelson), to specifically target viral misinformation circulating on these platforms, where the rising meme culture could also attract Snopes.com's attention and therefore explain a surge of digitally altered images (Ling et al., 2021; Wang et al., 2021).

2016 vs. 2020 US presidential election. We now compare misinformation types between the 2016 and 2020 elections. To filter for relevance, we constrain our analysis to fact-checks that (1) were published in the election years and (2) included the names of the presidential candidates and/or their running mates (*e.g.,* "Joe Biden" and "Kamala Harris"). This results in 2,586 fact-checks for the 2016 election and 2,436 fact-checks for 2020.

The prevalence of each misinformation type is shown in Figure 5. We observe that the relative ratios of many misinformation types are similar between the two elections, *e.g.,* legends and tales ■ and bogus scams ■, while the 2016 election has more hoaxes ■, satires ■, etc. The most prevalent type during both elections is fabricated and misleading content ■, next to conspiracy theories ■.

H1N1 vs. COVID-19. Finally, we compare misinformation types between the H1N1 pandemic in 2009 and the COVID-19 pandemic. For H1N1 related fact-checks, we search for keywords "flu", "influenza", and "H1N1" in fact-checks and constrain the publication date until the end of 2012.[14] For COVID-19 related fact-checks, we search for keywords "COVID-19" and "coronavirus", and only consider fact-checks published in 2019 or later, which results in 833 fact-checks for the H1N1 pandemic and 656 fact-checks for COVID-19.

The relative ratio of each misinformation type is also shown in Figure 5. We observe that the prevalence of some misinformation types are sig-

[13] Users can submit a topic to Snopes.com on its contact page (Snopes.com, 2021c), the results from which may affect Snopes.com's priorities.

[14] WHO declared an end to the global 2009 H1N1 pandemic on August 10, 2010, yet misinformation about H1N1 continues to spread (Sundaram et al., 2013), therefore we extend the time window by two more years.

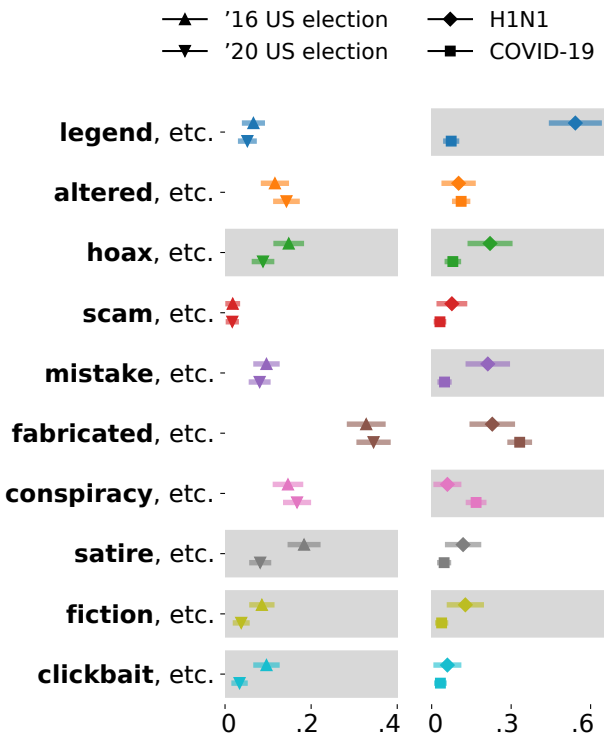

Figure 5: **Misinformation between notable events.** The most prevalent misinformation type for both US presidential elections is fabricated content, while the 2016 election has more hoaxes and satires. The H1N1 pandemic in 2009 has more legends and tales, while the COVID-19 pandemic attracts more conspiracy theories. (95% confidence intervals.)

nificantly different between two pandemics, *e.g.,* hoaxes ■, mistakes ■. Notably, the H1N1 pandemic has many more legends and tales ■, while COVID-19 has more conspiracy theories ■. The increased prevalence of COVID-19 related conspiracies aligns with recent work measuring the same phenomena (Uscinski et al., 2020; Jolley and Paterson, 2020), especially as the COVID-19 pandemic becomes increasingly politicized (Hart et al., 2020; Rothgerber et al., 2020; Weisel, 2021).

6 Discussion

In this section, we discuss limitations of our work and future directions, and finally conclude.

Limitations and future directions. We adopted a computational approach to investigate our research question, and this method inherently shares common limitations with observational studies, *e.g.,* prone to bias and confounding (Benson and Hartz, 2000). Specifically, our corpus contains fact-checks from Snopes.com, one of the most comprehensive fact-checking agencies in the US. Snopes.com covers a broader spectrum of topics than politics-focused fact-checkers (*e.g.,* PolitiFact.com, FactCheck.org),[15] and thus we argue that it covers a representative sample of misinformation within the US. However, Snopes.com may not be representative of the international misinformation ecosystem (Ahinkorah et al., 2020; Kaur et al., 2018; Fletcher et al., 2018). In the future, we hope that our method can help characterize misinformation comparatively on a global scale when more structured fact-checks become available.[16] Additionally, fact-checkers are time constrained, as thus the misinformation stories they cover tend to be high-profile. Therefore low-prevalence, long-tail misinformation stories may not be observed in our study. Understanding low-volume misinformation types may require a different collection of corpora other than fact-checks, *e.g.,* a cross-platform investigation on social media conversations (Wilson and Starbird, 2020; Abilov et al., 2021).

Lastly, the misinformation types we extract from our weakly supervised approach are not validated with ground-truth labels. This is largely due to the lack of empirical knowledge on misinformation types, and therefore we are unable to provide specific guidance to annotators. Although the clusters in Figure 3 provide straightforward structure of misinformation stories, in future work, we plan to leverage these results to construct annotation guidelines and obtain human-identified misinformation types for further analysis.

Conclusion. In this paper, we identify ten prevalent misinformation types with rationalized models on fact-checks and analyze their evolution over the last ten years and between notable events. We hope that this paper offers an empirical lens to the systematic understanding of fine-grained misinformation types, and complements existing work investigating the misinformation problem.

Acknowledgments

This research was supported in part by NSF grant IIS-1553088. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

[15] Also note that including these additional fact-checkers in the corpus would lead to oversampling of overlapping topics (*e.g.,* politics).

[16] Less-structured and under-represented fact-checks are difficult for computational modeling (Jiang et al., 2020a).

Ethical Considerations

This paper uses Snopes.com fact-checks to train and validate our models, and also includes several quotes and snippets of fact-checks.

We consider our case a *fair use* under the US[17] copyright law, which permits limited use of copyrighted material without the need for permission from the copyright holder.

According to 17 U.S.C. § 107, we discuss how our research abides the principles that are considered for a fair use judgment:

- Purpose and character of the use: we use fact-checks for noncommercial research purpose only, and additionally, using textual content for model training is considered to be transformative, *cf.* Authors Guild, Inc. v. Google Inc. (2013, 2015, 2016).
- Amount and substantiality: we present only snippets of fact-checks for illustrative purpose in our paper (*i.e.,* several quotes and snippets in text and figures), and only URLs to original fact-checks in our public dataset.
- Effect upon work's value: we do not identify any adverse impact our work may have on the potential market (*e.g.,* ads, memberships) of the copyright holder.

The end goal of our research aligns with that of Snopes.com, *i.e.,* to rebut misinformation and to restore credibility to the online information ecosystem. We hope the aggregated knowledge of fact-checks from our models can shed light on this road and be a helpful addition to the literature.

References

17 U.S.C. § 107. Limitations on exclusive rights: Fair use.

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 the International AAAI Conference on Web and Social Media (ICWSM)*.

Bright Opoku Ahinkorah, Edward Kwabena Ameyaw, John Elvis Hagan Jr, Abdul-Aziz Seidu, and Thomas Schack. 2020. Rising above misinformation or fake news in africa: Another strategy to control covid-19 spread. *Frontiers in Communication*.

Authors Guild, Inc. v. Google Inc. 2013. 954 f. supp. 2d 282 - dist. court, sd new york.

Authors Guild, Inc. v. Google Inc. 2015. 804 f. 3d 202 - court of appeals, 2nd circuit.

Authors Guild, Inc. v. Google Inc. 2016. 136 s. ct. 1658, 578 us 15, 194 l. ed. 2d 800 - supreme court.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Vian Bakir and Andrew McStay. 2018. Fake news and the economy of emotions: Problems, causes, solutions. *Digital journalism*.

Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. 2020. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*.

Kjell Benson and Arthur J Hartz. 2000. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL Interactive Presentation Sessions*.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sedona Chinn, P Sol Hart, and Stuart Soroka. 2020. Politicization and polarization in climate change news content, 1985-2017. *Science Communication*.

Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 113(3).

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 North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: a benchmark to evaluate rationalized nlp models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dan Evon. 2019. Is greta thunberg the 'highest paid activist'? *Snopes.com*.

[17] Where the authors and Snopes.com reside.

Richard Fletcher, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. 2018. Measuring the reach of "fake news" and online disinformation in europe. *Reuters institute factsheet*.

Vinny Green and David Mikkelson. A message to our community regarding the facebook fact-checking partnership. *Snopes.com*.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. *Science*.

P Sol Hart, Sedona Chinn, and Stuart Soroka. 2020. Politicization and polarization in covid-19 news coverage. *Science Communication*.

Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on youtube. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 4(CSCW).

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Shan Jiang, Simon Baumgartner, Abe Ittycheriah, and Cong Yu. 2020a. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proceedings of the Web Conference (WWW)*.

Shan Jiang, Miriam Metzger, Andrew Flanagin, and Christo Wilson. 2020b. Modeling and measuring expressed (dis)belief in (mis)information. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Shan Jiang, Ronald E Robertson, and Christo Wilson. 2019. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2020c. Reasoning about political bias in content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW).

Daniel Jolley and Jenny L Paterson. 2020. Pylons ablaze: Examining the role of 5g covid-19 conspiracy beliefs and support for violence. *British journal of social psychology*.

Kanchan Kaur, Shyam Nair, Yenni Kwok, Masato Kajimoto, Yvonne T Chua, Ma Labiste, Carol Soon, Hailey Jo, Lihyun Lin, Trieu Thanh Le, et al. 2018. Information disorder in asia and the pacific: Overview of misinformation ecosystem in australia, india, indonesia, japan, the philippines, singapore, south korea, taiwan, and vietnam. *Social Science Research Network (SSRN)*.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Chen Ling, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Dissecting the meme magic: Understanding indicators of virality in image memes. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 5(CSCW).

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Miriam Metzger, Andrew Flanagin, Paul Mena, Shan Jiang, and Christo Wilson. 2021. From dark to light: The many shades of sharing misinformation online. *Media and Communication*, 9(1).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Workshop Proceedings of the International Conference on Learning Representations (ICLR Workshop)*.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Gordon Pennycook and David Rand. 2021. Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *Harvard Kennedy School Misinformation Review*.

Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2021. What makes people join conspiracy communities? role of social factors in conspiracy engagement. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 4(CSCW).

Poynter. 2018. Verified signatories of the ifcn code of principles.

Radim Rehurek and Petr Sojka. 2011. Gensim - statistical semantics in python. *Gensim.org*.

Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW).

Hank Rothgerber, Thomas Wilson, Davis Whaley, Daniel L Rosenfeld, Michael Humphrey, Allie Moore, and Allison Bihl. 2020. Politicizing the covid-19 pandemic: ideological differences in adherence to social distancing. *PsyArXiv*.

Mattia Samory and Tanushree Mitra. 2018. 'the government spies using our webcams' the language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW).

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*.

Snopes.com. 2021a. Fact check ratings.

Snopes.com. 2021b. How does snopes decide what to write about?

Snopes.com. 2021c. Submit a topic.

Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 3(CSCW).

Maria E Sundaram, David L McClure, Jeffrey J VanWormer, Thomas C Friedrich, Jennifer K Meece, and Edward A Belongia. 2013. Influenza vaccination is not associated with detection of noninfluenza respiratory viruses in seasonal studies of influenza vaccine effectiveness. *Clinical infectious diseases*.

Joseph E Uscinski, Adam M Enders, Casey Klofstad, Michelle Seelig, John Funchion, Caleb Everett, Stefan Wuchty, Kamal Premaratne, and Manohar Murthi. 2020. Why do people believe covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Nguyen Vo and Kyumin Lee. 2018. The rise of guardians: Fact-checking url recommendation to combat fake news. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Nguyen Vo and Kyumin Lee. 2020. Where are the facts? searching for fact-checked information to alleviate the spread of fake news. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ellen M Voorhees. 1986. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*.

Yuping Wang, Fatemeh Tamahsbi, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, David Magerman, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the use of fauxtography on social media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Claire Wardle. 2017. Fake news. it's complicated. *First Draft News*.

Ori Weisel. 2021. Vaccination as a social contract: The case of covid-19 and us political partisanship. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 118(13).

Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology Innovation Management Review*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*.

Tom Wilson and Kate Starbird. 2020. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review*.

Lianwei Wu, Yuan Rao, Haolin Jin, Ambreen Nazir, and Ling Sun. 2019. Different absorption from the same sharing: Sifted multi-task learning for fake news detection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the Web Conference (WWW)*.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the ACM Internet Measurement Conference (IMC)*.

Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the ACM Web Science Conference (WebSci)*.

A Implementation Details

In this section, we discuss additional implementation details that we omitted in the main paper.

Loss functions. For the predictive loss $L_y(\hat{\boldsymbol{y}}, \boldsymbol{y})$, we use a common cross entropy loss function.

For the rationale regularization loss $L_z(\boldsymbol{z})$, we introduced it as a single item in the main paper for simplicity, but it actually contains two parts as implemented by Yu et al. (2019). The first part is to encourage conciseness:

$$L_{zk}(\boldsymbol{z}) = \max\left\{\sum_i \boldsymbol{z}^i - k, 0\right\},$$

where $\sum_i \boldsymbol{z}^i$ represents the number of selected tokens, and k is a hyperparameter defining a loss-free upper-bound for it. The second part is to encourage contiguity:

$$L_{zl}(\boldsymbol{z}) = \max\left\{\sum_i \left|\boldsymbol{z}^i - \boldsymbol{z}^{i-1}\right| - l, 0\right\},$$

where $\boldsymbol{z}^i - \boldsymbol{z}^{i-1}$ denotes a transition between $\boldsymbol{z}^i = 0$ and $\boldsymbol{z}^{i-1} = 1$ or vice versa, therefore $\sum_i \left|\boldsymbol{z}^i - \boldsymbol{z}^{i-1}\right|$ represents the number of rationale phrases, and l is another hyperparameter defining a loss-free upper-bound for it.

Combining these two parts together, we can further specify $\lambda_z L_z(\boldsymbol{z})$ as $\lambda_{zk} L_{zk}(\boldsymbol{z}) + \lambda_{zl} L_{zl}(\boldsymbol{z})$.

For domain knowledge weak supervision, we define $L_d(\boldsymbol{z}, \boldsymbol{z_d})$ as:

$$L_d(\boldsymbol{z}, \boldsymbol{z_d}) = -\sum_i \boldsymbol{z}^i \boldsymbol{z}^i_{\boldsymbol{d}},$$

which decreases loss by 1 if both $\boldsymbol{z}^i = 1$ and $\boldsymbol{z}^i_{\boldsymbol{d}} = 1$, *i.e.,* selecting a token in the domain knowledge vocabulary V_d, and has no effect on the loss otherwise. Similarly, we define $L_d(\boldsymbol{s}, \boldsymbol{z_d})$ as:

$$L_d(\boldsymbol{s}, \boldsymbol{z_d}) = -\sum_i \boldsymbol{s}^i \boldsymbol{z}^i_{\boldsymbol{d}},$$

which decreases loss by $\boldsymbol{s}^i$ if $\boldsymbol{z}^i_{\boldsymbol{d}} = 1$, and has no effect on the loss if $\boldsymbol{z}^i_{\boldsymbol{d}} = 0$. This encourages the training to increase the importance score $\boldsymbol{s}^i$ on domain knowledge to reduce the loss.

With this implementation, there are five hyperparameters to search for the hard rationalization method: λ_{zk}, k, λ_{zl}, l and λ_d, and only one hyperparameter to search for the soft rationalization method: λ_d.

Module cells. Each module in soft and hard rationalization methods can be implemented with different neural cells. Here, we consider two common types of choices: RNN cells, *e.g.,* LSTM, and transformer cells (Vaswani et al., 2017), *e.g.,* BERT (Devlin et al., 2019).

For hard rationalization, the rationale selection process is actively regularized by $L_z(\boldsymbol{z})$, therefore we simply choose the cell type that optimizes $\mathrm{F}_1(\boldsymbol{y})$ on dev sets, *i.e.,* transformers.

For soft rationalization, the rationale selection process is based on passively generated importance scores (*i.e.,* attention), therefore the inherent behavioral difference between RNN and transformer cells would significantly impact our choice.

	Test set evaluation					Dev set evaluation				
	Pr(z)	%(z)	$F_1(y)$	Ac(y)	%(x)	Pr(z)	%(z)	$F_1(y)$	Ac(y)	%(x)
Movie reviews (Zaidan et al., 2007)										
Hard rationalization	0.37	2.7%	0.72	0.72	2.7%	0.12	3.2%	0.71	0.71	3.5%
w/ Domain knowledge	0.38	3.7%	0.72	0.72	3.7%	0.14	3.9%	0.71	0.71	4.2%
w/o Rationale regu.	0.31	99.9%	**0.92**	**0.92**	99.9%	0.08	99.9%	**0.91**	**0.91**	99.9%
w/ Adversarial comp.	0.33	2.5%	0.70	0.70	2.5%	0.13	4.1%	0.70	0.70	3.7%
Soft rationalization	0.58	3.7%	0.91	0.91	100%	0.30	3.9%	0.90	0.90	100%
w/ Domain knowledge	**0.62**	3.7%	**0.92**	**0.92**	100%	**0.33**	3.9%	**0.91**	**0.91**	100%
Personal attacks (Carton et al., 2018)										
Hard rationalization	0.17	32.5%	0.73	0.73	32.5%	0.19	30.2%	0.74	0.74	30.2%
w/ Domain knowledge	0.22	16.9%	0.73	0.73	16.9%	0.23	15.7%	0.74	0.74	15.8%
w/o Rationale regu.	0.19	99.9%	**0.82**	**0.82**	99.9%	0.20	99.9%	0.84	0.84	99.9%
w/ Adversarial comp.	0.22	14.9%	0.75	0.75	14.9%	0.23	15.2%	0.76	0.76	15.2%
Soft rationalization	0.35	16.9%	**0.82**	**0.82**	100%	0.37	15.7%	0.84	0.84	100%
w/ Domain knowledge	**0.39**	16.9%	**0.82**	**0.82**	100%	**0.40**	15.7%	**0.85**	**0.85**	100%
Fact-checks										
Soft rationalization	-	-	0.74	0.83	100%	-	-	0.72	0.83	100%
w/ Domain knowledge	-	-	**0.75**	**0.85**	100%	-	-	**0.73**	**0.85**	100%

Table 2: **Evaluation results on both test and dev sets for hard and soft rationalization methods.** An additional accuracy metric Ac(y) is included, as well as results for the fact-checks dataset. The results on dev sets align with our findings on test sets in the main paper.

In our experiments, we observe that transformer cells often assign strong importance to a single token, but assign near zero weights to its neighboring tokens (possibly as a result of its multi-head attention mechanism), while RNN cells assign strong importance to a single token, but also some residue, fading weights to its neighboring tokens.

Consider the following example, which shows the distribution of importance scores generated by transformer cells, with **darker** text representing higher importance scores and lighter text scoring near zero. In the following example, only the token **conspiracy** is selected as rationale:

> "...Furthermore, claims that COVID-19 was "manufactured," or that it "escaped from" this Chinese lab, are nothing more than baseless **conspiracy** theories..."

In contrast, the following example shows the distribution of importance scores generated by RNN cells for the same snippet, *i.e.,* the token **conspiracy** has the strongest importance score, but its neighboring tokens are also assigned some weight above the threshold, and therefore the phrase **baseless conspiracy theories** is selected as rationale:

> "...Furthermore, claims that COVID-19 was "manufactured," or that it "escaped from" this Chinese lab, are nothing more than baseless **conspiracy** theories..."

As we prefer to obtain phrases (*i.e.,* one or more tokens) for rationales, we choose between RNN cells. After optimizing $F_1(y)$ on dev set, we choose bidirectional LSTM initialized with GloVe embeddings (Pennington et al., 2014) for the soft rationalization method.

Hyperparameters. As discussed in the paper, we optimize hyperparameters for $F_1(y)$ on the dev sets.

Since the size of dev sets is relatively small in our experiments, a rigorous grid search for hyperparameters might overfit to several instances in the dev set, therefore we tune the hyperparameters manually starting from the hyperparameters released by (Yu et al., 2019) and (Carton et al., 2018).

For **movie reviews** (Zaidan et al., 2007), the best-performing model for hard rationalization uses $\lambda_{zk} = 5.0$, $k = 240$, $\lambda_{zl} = 5.0$, $l = 10$, and $\lambda_d = 8.0$ with domain knowledge as weak supervision, and the best-performing model for soft rationalization uses $\lambda_d = 0.5$.

For **personal attacks** (Carton et al., 2018), the best-performing model for hard rationalization uses $\lambda_{zk} = 5.0$, $k = 7$, $\lambda_{zl} = 5.0$, $l = 1$, and $\lambda_d = 10.0$ with domain knowledge as weak supervision, and the best-performing model for soft rationalization uses $\lambda_d = 0.5$.

For **fact-checks**, the best-performing model for soft rationalization uses $\lambda_d = 1.0$.

Domain knowledge for fact-checks. V_d contains the following words, in which the first 5 are from Wardle (2017) and the remaining 7 are

from Snopes.com (2021a):

> "fabricated, manipulated, imposter, misleading, parody, satire, unproven, outdated, scam, legend, miscaptioned, misattributed."

B Additional Results

In this section, we record additional results from our experiments that we omitted in the main paper.

Validation performance. The evaluation results for all our experiments on both test and dev sets are reported in Table 2. We also include accuracy metric $\text{Ac}(\boldsymbol{y})$ in the table[18], and the evaluation results for fact-checks. Note that evaluation for z is empty for fact-checks, since there are no ground-truth rationales. As shown in Table 2, the results on dev sets align with our findings on test sets discussed in the main paper.

Model size, computing machine and runtime. The number of parameters is 325K for hard rationalization models, and 967K for soft rationalization models. All experiments were conducted on a 12GB Nvidia Titan X GPU node, and finished training within an hour per experiment.

C Rationale Examples

In this section, we list additional examples of extracted rationales for ten identified misinformation types.

For urban legends and tales ■:

> "...the 1930 Colette short story La Chienne (The Bitch) has become an **urban legend** in that its plot is often now related as a string of events that..."

For altered or doctored images ■:

> "...magazine covers of "highest paid" people. These **doctored images** have featured celebrities such as John Legend, Chuck Norris, Bob Dylan, Susan Boyle, and..."

For hoaxes and pranks ■:

> "...This meme is a **hoax**. Nobody is (or was) licking toilets as a form of protest against Donald Trump. The images shown in the meme were taken from..."

For bogus scams ■:

> "...In October 2019, we came across a decidedly bizarre version of **the scam**. This time, Nigerian astronaut Abacha Tunde was reportedly stuck in space and..."

For mistakes and errors ■:

> "...noted that reports of missing children (which are typically resolved quickly) are often **mistakenly** confused by the public with relatively rare instances of..."

For fabricated content ■:

> "...The Neon Nettle report was "unusual" because it was **completely fabricated**: Bono said nothing during his Rolling Stone interview about "colluding with elites"..."

For baseless conspiracies ■:

> "...Furthermore, claims that COVID-19 was "manufactured," or that it "escaped from" this Chinese lab, are nothing more than **baseless conspiracy theories**..."

For satires and parodies ■:

> "...This item was not a factual recounting of real-life events. The article originated with a website that describes its output as being **humorous** or **satirical** in nature..."

For fictitious content ■:

> "...However, both of these shocking quotes, along with the rest of article in which they are found, are **completely fictitious**. As the name of the web site implies..."

For sensational clickbait ■:

> "...And Breitbart regurgitated some of the pictures as viral **clickbait** under the headline "Armed Black Panthers Lobby for Democrat Gubernatorial Candidate Stacey Abrams"..."

[18] Our public dataset has balanced positive and negative labels therefore $\text{Ac}(\boldsymbol{y}) = \text{F}_1(\boldsymbol{y})$.