ClaimDiff: Comparing and Contrasting Claims on Contentious Issues

Miyoung Ko[a,⋆] **Ingyu Seong**[b,†] **Hwaran Lee**[c]
Joonsuk Park[c,d] **Minsuk Chang**[c,‡] **Minjoon Seo**[a]
[a]KAIST [b]Korea University [c] NAVER AI Lab [d] University of Richmond
{miyoungko, minjoon}@kaist.ac.kr
dlssrb7777@korea.ac.kr park@joonsuk.org
{hwaran.lee, minsuk.chang}@navercorp.com

Abstract

With the growing importance of detecting misinformation, many studies have focused on verifying factual claims by retrieving evidence. However, canonical fact verification tasks do not apply to catching subtle differences in factually consistent claims, which might still bias the readers, especially on contentious political or economic issues. Our underlying assumption is that among the trusted sources, one's argument is not necessarily more true than the other, requiring *comparison* rather than *verification*. In this study, we propose ClaimDiff, a novel dataset that primarily focuses on *comparing* the nuance between claim pairs. In ClaimDiff, we provide 2,941 annotated claim pairs from 268 news articles. We observe that while humans are capable of detecting the nuances between claims, strong baselines struggle to detect them, showing over a 19% absolute gap with the humans. We hope this initial study could help readers to gain an unbiased grasp of contentious issues through machine-aided comparison.

1 Introduction

With an ever-increasing amount of textual information on the web, many researchers have focused on detecting misinformation from diverse sources, such as fake news (Potthast et al., 2018; Nguyen et al., 2020) and rumor tweets (Zubiaga et al., 2016; Kochkina et al., 2018). In particular, fact verification has become a popular task due to its utility and the availability of reliable datasets such as FEVER (Thorne et al., 2018; Aly et al., 2021).

For contentious issues, however, fact verification alone is not sufficient; comparing and contrasting claims from the opposing sides are necessary to gain an unbiased understanding of the issue. Figure 1 presents two articles reporting on the treatment of long COVID – long-term physical and mental symptoms that can occur after COVID-19 infection. Although both articles are published in the same week, their perspectives on long COVID are quite different; article A downplays the risk of long COVID stating that it has become less common in the recent COVID variants, while article B expresses concerns about the increase in the number of people suffering from long COVID. In this way, even articles from trusted sources may provide biased views about an issue.

In this paper, we present ClaimDiff, a novel dataset consisting of 2,941 claim pairs extracted from 268 news articles on 134 contentious issues.[1] ClaimDiff comes in two variations—ClaimDiff-S and ClaimDiff-W—each consisting of labels targeting a different relation: determining whether a claim **S**trengthens and **W**eakens another claim, respectively. For instance, the two claims in Figure 1 weaken each other, providing inconsistent perspectives surrounding the undisputed factual information — the human body has a natural ability to heal from long COVID. More specifically, claim A argues that medical treatments should be geared toward supporting the natural ability to heal, while claim B calls for more active interventions for fast recovery. ClaimDiff focuses on recognizing such relations between two claims on an issue; this is distinguished from existing tasks such as fact verification — verifying the veracity of a claim using evidence text (Vlachos and Riedel, 2014; Thorne et al., 2018) — and stance detection — identifying the stance of a claim toward a topic of interest (Mohammad et al., 2016; Derczynski et al.).

We also demonstrate the efficacy of ClaimDiff on two tasks it supports — relation classification and rationale extraction — and an extended ap-

⋆This work was done during internship in NAVER AI Lab.
†Is now at Samsung Research.
‡Is now at Google.

[1]The articles are collected from allsides.com, licensed under a CC BY-NC 4.0 license.

Findings of the Association for Computational Linguistics: ACL 2023, pages 4711–4731
July 9-14, 2023

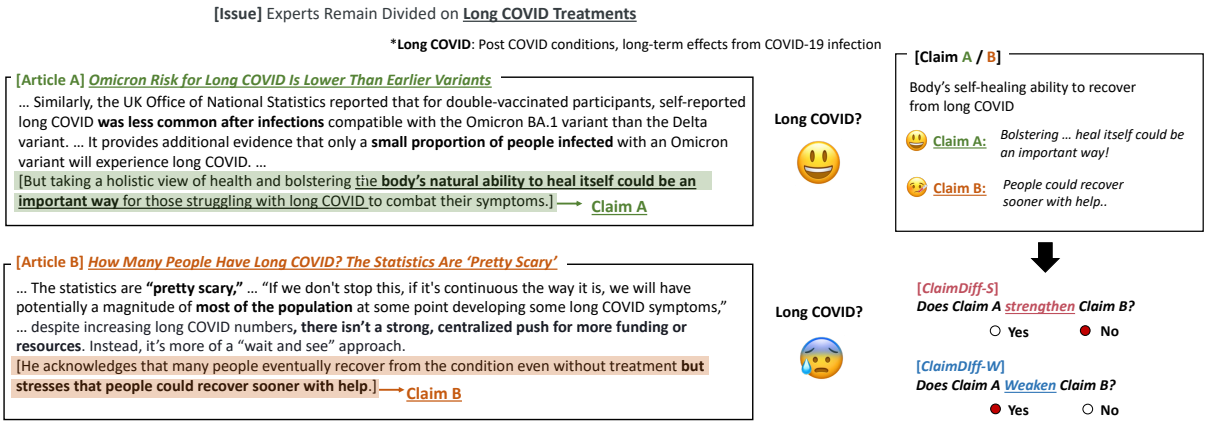


Figure 1: Two articles are reporting on long COVID with opposite perspectives: (A) Recent variants have less risk for long COVID, while (B) Statistics for long COVID is *pretty scary*. This might lead readers to have different understandings of long COVID. ClaimDiff targets the "strengthening" and "weakening" relationships between two claims on the same issue. Although both claims are about self-healing ability to overcome long COVID, the nuances are different, having "weakening" relation. The rationale for 'A weakens B' relation is underlined in Claim A.

plication — document-level ClaimDiff. Relation classification experiments tests the ability to identify strengthen and weaken relations. Rationale extraction tests the ability to find the rationale for a given relation in extractive and generative scenarios. Lastly, although ClaimDiff only provides sentence-level annotations, models trained on it allow document-level analyses. In this document-level ClaimDiff, we test the potential of analyzing points of agreement and disagreement in documents.

The main contributions of our work are:

- We present ClaimDiff, a novel dataset of claim pairs on contentious issues. Each pair comes with strengthen and weaken relation labels, as well as a rationale for choosing the label.
- We present competitive baselines for the dataset, leveraging finetuning, parameter-efficient finetuning, and zero-shot approaches, along with human performance.
- We showcase how models trained on ClaimDiff can be used to compare documents on an issue, indicating specific points of agreement and dispute, in document-level ClaimDiff.

2 Related Works

Dealing with Misinformation Detecting and avoiding misinformation received intense interest with a massive amount of information on the web. Existing works introduced benchmarks with a broad spectrum of sources, including rumors in social media (Potthast et al., 2018; Kochkina et al., 2018) and fake news (Zubiaga et al., 2016). Other researchers focus on dealing with an exploding amount of misinformation on global events, such as the COVID-19 pandemic (Saakyan et al., 2021; Jiang et al., 2021; Alam et al., 2021; Weinzierl and Harabagiu, 2022). While many existing benchmarks aim to detect less reliable information on unverified sources, our work targets subtle differences in trusted sources.

Claim Verification Claim verification verifies the factuality of the target sentence with respect to a reliable truth from verified sources. Automatic verification shows remarkable progress with the introduction of rich claim verification datasets (Vlachos and Riedel, 2014; Thorne et al., 2018; Hanselowski et al., 2019; Aly et al., 2021; Khan et al., 2022). Existing claim verification datasets introduce many variants, including a shift in domains and languages of claims; claims from political sources (Wang, 2017; Garimella et al., 2018), scientific claims (Wadden et al., 2020), climate change-related claims (Leippold and Diggelmann, 2020), Arabic claims (Baly et al., 2018; Alhindi et al., 2021), and Danish claims (Nørregaard and Derczynski, 2021). Our assumption is different from claim verification in that one claim is not necessarily more true than the other. As a result, ClaimDiff focues on comparison between claims rather than verification.

Stance Detection Stance detection aims to predict the stance of a claim toward a specific topic between agreeing or opposing perspectives. Mohammad et al. (2016) propose SemEval challenges to

predict the stance of tweets toward target keywords. Derczynski et al. further presents a sub-challenge to detect the stance of corresponding threads of rumor tweets. Other benchmarks are introduced with diverse challenges, such as claim-based stance detection (Ferreira and Vlachos, 2016; Bar-Haim et al., 2017), stance detection with evidence (Chen et al., 2019), and stance detection over political domains (Li et al., 2021).

3 ClaimDiff

In this section, we formally define the tasks supported by ClaimDiff, describes how the dataset is constructed, and provide the statistics and analysis of the resulting dataset.

3.1 Task Description

ClaimDiff comes in two variations—ClaimDiff-S targeting strengthen relations, and ClaimDiff-W targeting weaken relations. Both variations of the dataset were designed to support the following tasks.

Relation Classification Relation classification aims to determine if claims from two different documents are in a relation: strengthen for ClaimDiff-S, and weaken for ClaimDiff-W. For instance, as shown in Figure 1, claim A and claim B are both about the treatment of long COVID. Claim A and B exhibit opposing positions for the body's natural recovery, respectively. We want to classify this case into *weakens* as claim A weakens claim B. Although, in this case, claim B weakens A as well, note that the relationship is not guaranteed to be symmetric in general.

More formally, for ClaimDiff-S, given a claim pair (c_1, c_2), the objective is to return *true* if c_1 strengthens c_2, and *false*, otherwise. For ClaimDiff-W, given a claim pair (c_1, c_2), the objective is to return *true* if c_1 strengthens c_2, and *false*, otherwise. Note that for any given claim pair, the answer cannot be *true* for both variations of ClaimDiff.

Rationale Extraction Rationale extraction aims to extract phrases from a claim in a relation: strengthen for ClaimDiff-S, and weaken for ClaimDiff-W. For instance, in Figure 1, a rationale for ClaimDiff-W is *'body's natural ability to heal itself could be an important'*, which weakens claim B about the view that more active interventions for fast recovery.

More formally, given a claim pair (c_1, c_2) in a relation, the objective is to extract phrases in c_1 that provide a rationale for relation. Note that a single pair of claims can have multiple rationales.

3.2 Constructing ClaimDiff Dataset

Raw Data Collection We first collect a group of news articles from AllSides[2] headlines. Allsides provides a balanced search of news with all sides of the political spectrum. More specifically, in AllSide headlines, there are groups of news articles about the same issues from different media press. The article groups have a broad political spectrum; each belongs to one of (left, center, and right) political stances. We crawl the headline pages uploaded from 2012-06-01 to 2021-11-21, covering more than 180 media sources. We choose two articles with left and right labels if possible.

To construct the claim pairs, we filter the non-overlapping sentence pairs from the article pair. As most claim pairs do not provide overlapping contents, a large proportion of pairs are filtered out. We apply an additional filtering process using Amazon Mechanical Turk (MTurk) [3] to collect the overlapping claim pairs. Each worker is asked to answer the question, *'Does the target sentence overlap with a given sentence?'*, where each sentence is extracted from two different articles. Given a single example, three workers made a response. We collected the pairs if at least two workers answered that the given pairs are "overlapping". The details of the filtering process are shown in Appendix A.

Annotation Process After the filtering process, we conducted an annotation process with 15 in-house expert annotators to obtain the final data. Given a pair of claims, the annotators were requested to determine the stance among *strengthen*, *weaken* and *no effect*. If they choose *strengthen* or *weaken*, the annotators had to select the phrases from the claim that strengthen or weaken the other claim. The overall interface for the data collection process is shown in Figure 4.

For each single claim pair, three to five participants submitted their responses. We collect the responses and convert the relation options to scaled values. We first consider *strengthen* as 1, *weaken* as -1, and *no effect* as 0 and average the choices after conversion. We filter out the pairs with absolute average values between (0, 0.5), which means the

[2] https://www.allsides.com/unbiased-balanced-news/
[3] https://www.mturk.com/

	Train	Test	Test-doc
Pairs	1,857	1,084	3,173
Issues	90	44	44
Articles	180	88	88
Rationales	1,484	852	852
% Strengthen	69.31%	56.64%	19.35%
% Weaken	10.61%	21.96%	7.50%

Table 1: Statistics of ClaimDiff dataset. Test-doc indicates the raw test dataset over the whole article, including non-overlapping claims.

relations are ambiguous.[4] The pairs with positive values are mapped into strengthening claims (1 for ClaimDiff-S), and the pairs with negative values are mapped into weakening claims (1 for ClaimDiff-W). If the resulting values are exactly 0, the claims become 0 for both ClaimDiff-S and ClaimDiff-W. We measure the average inner-annotation agreement by Krippendorff's alpha (Hayes and Krippendorff, 2007). The scores are 0.46 and 0.47 for ClaimDiff-S and ClaimDiff-W, respectively.

Constructing Test-doc Dataset We aim to build an application for understanding contentious issues with diverse views on a fine-grained level. Rather than classifying an article in a single label (i.e., containing left or right political bias), ClaimDiff enables a comparison between two articles in a sentence-wise manner. However, this requires a further extension of ClaimDiff from sentence pairs to document pairs, having significantly different distributions. To be coherent with the real-world distribution over whole articles, we provide an additional test dataset, *test-doc*, following the distribution over article pairs. Test-doc can be considered as an unfiltered test dataset, resulting in a high ratio of non-overlapping pairs. We collect the non-overlapping claim pairs of articles in the test dataset that are obtained from the previous filtering step. Over 70% of the claim pairs from article pairs are not overlapped, resulting in a highly skewed distribution. We combine filtered-out claim pairs with a standard test dataset to construct the final test-doc dataset.

3.3 Dataset Statistics

We extract the pairs from a group of articles sharing the same topic, published by multiple presses. Our final dataset contains articles from 47 presses. The top-3 presses with the highest appearance are *Fox News*, *CNN*, and *Washington Times*. We further analyze the topic diversity of our dataset by collecting the tag information. *Tag* is the list of words representing the topic provided in Allsides headlines. For instance, tags for a topic, *Supreme Court Sides With Google in Copyright Dispute Case*, are *Supreme Court*, *Copyright*, *Google*, and *Oracle*. The number of unique tags is 276, while each topic includes an average of 3.6 tags. The most common issues are *Elections*, *Donald Trump*, and *Coronavirus*. The overall lists of presses and tags are presented in Appendix D.

Table 1 presents the overall statistics of ClaimDiff. ClaimDiff dataset provides 2,941 examples, extracted from 268 articles with 134 issues. Note that ClaimDiff-S and ClaimDiff-W consists of the same claim pairs with different labels. Since non-overlapping claim pairs do not provide rationales, the number of rationales is less than the overall claim pairs. Each pair contains an average of 1.4 rationales with an average length of 13.2 tokens.[5] In the train and test dataset, "strengthening" pairs are available to be found with more than 50% appearance. Finding the "weakening" claim pairs is more challenging, resulting in 7.50% weakening examples in the test-doc environment.

3.4 Dataset Analysis

This part analyzes claim pairs in ClaimDiff with respect to the class label. We first provide the subjectivity analysis over claims with positive labels in ClaimDiff-S and ClaimDiff-W. ClaimDiff includes both subjective and objective claims, indicating the proposed task is designed to predict a more general relation between each claim pair. We further present prediction results of the natural language inference (NLI) and fact verification (FEVER) models. The results indicate that models trained on the datasets are not suitable for understanding the nuances.[6]

Subjectivity Analysis To analyze the subjectivity of claims, we randomly sample 50 "strengthening" examples from ClaimDiff-S and 50 "weakening" examples from ClaimDiff-W test data. We manually label each claim in a pair with *subjective* or *objective*. Following Wiebe and Riloff (2005), we distinguished subjective and objective claims based on whether each claim includes at least one private state – opinions, evaluations, emotions, and speculations. We found that "strengthening" pairs

[4] Unfiltered data with average response scores are also publicly available.

[5] We use spaCy tokenizer for tokenizing the rationales.

[6] Check the Appendix E for additional information.

have a high proportion of objective claims on both claim A and claim B (A: 72%, B: 80%). However, "weakening" examples include more than 40% of subjective claims (A: 44%, B: 48%), indicating more diverse patterns in "weakening" relations. We expect that ClaimDiff-W to be more challenging not only because of the skewed distribution but also because of the more diverse composition of claims.

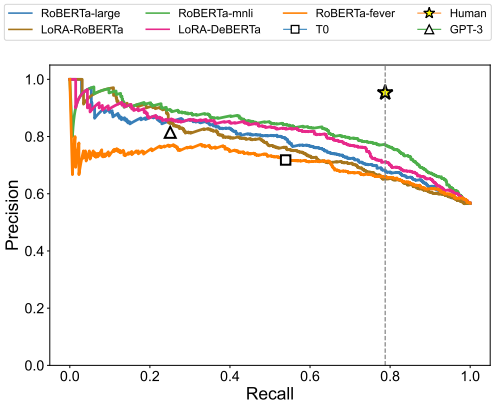
Prediction Results of NLI / FEVER Model To compare the ClaimDiff with previous sentence pair classification tasks, we analyze prediction results of transformer-based models trained on NLI and FEVER. We use RoBERTa-large trained on MNLI (Williams et al., 2018) and FEVER (Thorne et al., 2018). Each model yielded 90.2 (MNLI) and 75.6 (FEVER) F1 scores, respectively. Among 614 "strengthening" and 239 "weakening" pairs, the MNLI model predicts 561 and 203 pairs as *neutral*. FEVER model predicts 532 (strengthen) and 226 (weaken) examples as *not enough info*. These failures might come from multiple reasons, including the domain shift in claims and the different goals of each task. However, we observe that "weakening" pairs contain a slightly higher ratio of *contradiction* (strengthen: 5.7% vs. weaken: 15%) and 0 *entailment* with the MNLI model. "Strengthening" pairs show the difference in FEVER, containing more *support* (FEVER, 9.0% vs. 1.6%) examples. MNLI and FEVER models might be able to distinguish weakening and strengthening examples from others, which can work like the prior knowledge for solving ClaimDiff in Section 4.

4 Task 1 - Relation Classification

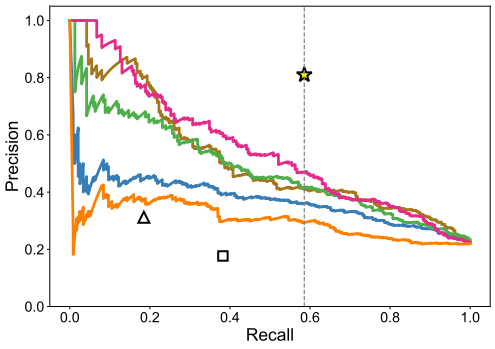
4.1 Experimental Setup

In this section, We present the baselines with different learning strategies: (1) finetuning, (2) parameter-efficient finetuning, and (3) zero-shot baselines. Implementation details of each model are described in Appendix B.

Finetuning Baselines We finetune pre-trained language models for the sentence classification tasks. Each model takes a pair of claims as inputs and predicts whether the first claim strengthens/weakens the other. We finetune models with a weighted loss function, where the class weight is determined by the label distribution.[7] We construct the development data by sampling 30% of



(b) ClaimDiff-S



(c) ClaimDiff-W

Figure 2: Precision-Recall curve of the baselines. Zero-shot baselines are represented as points as it is not feasible to control threshold for zero-shot predictions.

issues from training data. Following the test environment, issues in development data are exclusive. We train RoBERTa-base and RoBERTa-large (Liu et al., 2019) on ClaimDiff-S and ClaimDiff-W, respectively. We further present RoBERTa (FEVER) and RoBERTa (MNLI), RoBERTa-large initialized on MNLI and FEVER.[8]

Parameter-efficient finetuning As the number of training examples is not large enough, we further explore the parameter-efficient finetuning methods. Following Hu et al. (2022), we apply low-rank adaptation (LoRA) on pre-trained language models, which only finetune the task-specific low-rank matrices. We follow the same procedure as finetuning baselines for model selection. We compare the effect of LoRA on RoBERTa-large (355M) and DeBERTa-XXL (1.5B) (He et al., 2021).

Zero-shot Baselinse We present the zero-shot performance of large language models, T0 (Wei et al., 2021) and GPT-3 (Brown et al., 2020; Ouyang et al., 2022). Both models get a test pair

[7]Training without the weight is not possible due to the extremely skewed distribution of ClaimDiff-W, resulting 0 F1.

[8]We use the same models as in Section 3.4 for initialization.

Model	# Trainable Param.	*ClaimDiff-S* AUROC	F1	Precision	Recall	*ClaimDiff-W* AUROC	F1	Precision	Recall
finetuning									
RoBERTa-base	125M /125M	0.7160	71.82	68.97	74.92	0.6350	38.04	37.05	39.08
RoBERTa-large	355M / 355M	0.7085	71.37	67.49	75.73	0.7137	44.89	35.53	60.92
RoBERTa (FEVER)	355M /355M	0.6841	72.61	64.21	83.55	0.6056	36.77	24.29	**75.63**
RoBERTa (MNLI)	355M / 355M	**0.7976**	**77.86**	70.80	**86.48**	0.7481	45.33	48.11	42.86
LoRA									
RoBERTa-large	0.8M / 355M	0.5194	61.83	55.34	70.03	0.7633	42.13	49.71	36.55
DeBERTa-XXL	4.7M / 1.5B	0.7668	76.34	69.72	84.36	**0.7685**	**49.07**	**54.64**	44.54
zero-shot									
T0	0 / 11B	-*	61.58	71.80	53.91	-*	24.17	17.67	38.24
GPT-3	0 / 175B	-*	38.36	**81.48**	25.08	-*	23.22	31.21	18.49
Human Evaluation	-	-	86.27	95.34	78.77	-	68.29	81.01	58.56

Table 2: Performance on test dataset. *# Trainable Param.* represents the number of trainable parameters over the number of model parameters. Human indicates the human performance evaluated on test dataset. (*AUROC is not defined as threshold is not applicable for zero-shot generation.)

following the prompt format and generate the answer directly. We consider the class token with maximum probability as the prediction results. The examples of the zero-shot prompts are presented in Appendix B.

4.2 Evaluation on ClaimDiff Test

Human Evaluation Following Rajpurkar et al. (2016), we evaluate human performance on ClaimDiff based on the human annotation results. As each example has at least three responses, we randomly sample one response as the human prediction. We obtain ground-truth labels using the remainder following the same procedures as in Section 3.2. The resulting human performance is shown in Table 2. Humans are capable of detecting both types of relation, resulting in a significant gap between human and model performance. Even for more challenging ClaimDiff-W, humans can detect more than half of the weakening nuances while maintaining 81.01% precision.

Main Results Figure 2 shows the precision-recall curve of our baselines. Compared to the vanilla RoBERTa-large, initialization with FEVER worsen the performance while MNLI gives a significant gain in both ClaimDiff-S and ClaimDiff-W. Parameter-efficient finetuning (LoRA) is effective for ClaimDiff-W, even when using a same size model (RoBERTa-large). This is due to the small number of "weakening" examples, which makes finetuning the whole parameters more difficult. When recovering about 80% of "strengthening" examples (ClaimDiff-S), humans retain over 90% of precision, while the best working model retains 80%. The gap is more significant in ClaimDiff-W, showing about a 30% of difference in precision. Zero-shot baselines, T0 and GPT-3, are worse than finetuning RoBERTa-large in both tasks. We use prompts asking about a single relation ("support" / "weaken") for zero-shot baselines, while the actual relations in ClaimDiff are more complex, which results low coverage of zero-shot models (i.e., GPT-3 predicts only 17% of examples as "strengthening").

The overall evaluation results are presented in Table 2. Note that we report AUROC except for zero-shot baselines, as zero-shot generation does not require any threshold. Align with previous observations, initialization by MNLI boost the performance on both tasks, obtaining the best AUROC in ClaimDiff-S. LoRA becomes more effective when the number of examples is small. LoRA enables training the 1.5B size model (DeBERTa-XXL) with only the hundreds of "weakening" examples, obtaining the best AUROC and F1 in ClaimDiff-W.

Error Analysis To further understand the challenges in ClaimDiff, we investigate the errors of finetuned RoBERTa-large. We randomly sample 25 false negatives (i.e., relationships that the model failed to detect) from each ClaimDiff-S and ClaimDiff-W.[9] We manually categorize the errors and analyze them. The examples of each category and the ratio are provided in Appendix F. The errors in "strengthening" have relatively simple patterns (ex., 28% entailment or 20% coherent nuance). However, it is more difficult to capture these relationships in ClaimDiff as claims are collected from real-world news articles, resulting relatively low lexical overlap between two claims. On

[9]We also analyzed the false positives (FP) while the patterns of FP are too diverse to capture the common patterns.

the other hand, we find that patterns in ClaimDiff-W are more challenging; we suspect that this is because there are diverse ways to weaken one's argument. Finally, a common error (12%) in ClaimDiff-S and ClaimDiff-W is due to the need for context information or background knowledge to understand the claims.

5 Task 2 - Rationale Extraction

We perform rationale extraction over "strengthening" and "weakening" pairs, which have positive labels for ClaimDiff-S and ClaimDiff-W. Because of the low appearance of "weakening" pairs, training individual models for each task is challenging. Therefore, unlike relation classification, we train a single model to extract rationales from both "strengthening" and "weakening" pairs. We experiment with extractive and generative models. The evaluation is also conducted on combined sets of ClaimDiff-S and ClaimDiff-W.

5.1 Models

Extractive Model In this work, we present machine reading comprehension (MRC) models as the extractive baselines. As shown in Figure 1, rationales are found as the phrases existing in input claims. Extracting the phrases from a given text is similar to the previous MRC (Rajpurkar et al., 2016; Trischler et al., 2017). Following Devlin et al. (2019), we finetune pre-trained language models with the output layer that predicts the start and end positions of given rationales. We train RoBERTa-base and RoBERTa-large for rationale extraction.

Generative Model Generative baselines directly *generate* the rationales rather than extract it from input claims. Existing works (Narang et al., 2020; Lakhotia et al., 2021) show that generative models obtain a strong performance on rationale benchmark, ERASER (DeYoung et al., 2020). ERASER is designed to evaluate the reasoning ability of NLP models, containing 7 NLP tasks, including BoolQ (Clark et al., 2019), and Movie Reviews (Zaidan and Eisner, 2008). Following Narang et al. (2020), we finetune the T5 models to sequentially generate a list of rationales in a token-by-token fashion. We experiment with T5 and T5 with "strengthening" / "weakening" labels. The exact input formats are explained in Appendix B.

	Perplexity	TF1	IOU F1
RoBERTa-base	-	63.67	57.25
RoBERTa-large	-	63.49	55.99
T5-base	1.49	72.78	65.05
+ *class label*	1.45	72.70	63.95
T5-large	1.45	75.08	**67.74**
+ *class label*	1.46	**77.01**	**67.70**

Table 3: Rationale extraction performance measured on test set. Note that it is possible to measure perplexity only for generative baselines (T5).

5.2 Evaluation Metrics

Perplexity We report the per-token perplexity of rationales that measures how well the language model predicts the tokens in each rationale. Perplexity is defined as the exponentiated average negative log-likelihood of a sequence. Note that perplexity is only for the generative baselines.

Token F1 (TF1) Following Lakhotia et al. (2021), we compute the Token-level F1 between ground-truth rationales and generated rationales. TF1 measures the number of overlapping tokens between two rationales. Following DeYoung et al. (2020), we use spaCy tokenizer[10] to compute the F1 score.

Intersection over Union F1 (IOU F1) IOU F1, as used in DeYoung et al. (2020), computes the F1 on matched predictions. IOU F1 first checks whether predicted rationales match ground-truth rationales by calculating the intersection of union (IOU). IOU is computed as the number of overlapping tokens divided by the union of tokens. If IOU is larger than the threshold, the predicted explanation becomes a matched prediction. In this work, we set the threshold as 0.5.

5.3 Results

In Table 3, we show the results of finetuned RoBERTa and T5 trained on rationale extraction. Following DeYoung et al. (2020), we report extractive measures (TF1 and IOUF1), as the ground-truth rationales are extracted phrases from each claim. We further measure the generative score (perplexity) of output sequences for generative models. Note that we choose perplexity as the metric because rationales are in phrases rather than complete sentences. Although ClaimDiff contains multiple phrases as the ground-truth rationales, MRC models predict a single rationale for each claim pair. Since T5 is capable of generating multi-

[10] https://spacy.io/

	ClaimDiff-S				*ClaimDiff-W*			
	AUROC	F1	Precision	Recall	AUROC	F1	Precision	Recall
finetuning								
RoBERTa-base	0.6955	40.28	27.54	74.92	0.6303	19.70	13.17	39.08
RoBERTa-large	0.7137	38.93	25.00	**87.95**	0.6646	19.75	11.79	60.92
RoBERTa (FEVER)	0.7253	40.76	26.96	83.55	0.5981	15.35	8.54	**75.63**
RoBERTa (MNLI)	**0.7510**	41.27	27.11	86.48	0.7005	24.46	17.11	42.86
LoRA								
RoBERTa-large	0.6620	36.57	24.11	75.73	0.7282	24.17	18.05	36.55
DeBERTa-XXL	0.7266	40.09	26.29	84.36	**0.7289**	**26.70**	**19.06**	44.54
zero-shot								
T0	-	**41.77**	34.09	53.91	-	13.37	8.10	38.24
GPT-3	-	32.29	**45.29**	25.08	-	17.25	16.18	18.49

Table 4: Performance measured on test-doc split. The baselines are the same as those described in Section 4, but evaluated on a different test data.

ple phrases at once, even smaller T5-base obtains better performance than RoBERTa-large. T5-large consistently provides better results than T5-base regardless of whether the class labels are given or not. The injection of labels degrades the performance of T5-base, while the T5-large shows a slight improvement. The increasing number of parameters is also beneficial for incorporating additional label information.

6 Extension: Document-level ClaimDiff

Suppose we want to compare the articles with opposing views on contentious issues. For instance, there are two articles about a topic, *"Will Gas Prices Come Down Soon or Stay High?"*. One forecasts the *increase* in gas prices, while the other supports the prices have *already reached a peak*. A single stance label on the relation between the articles (i.e., whether one article supports or opposes the other) might not be enough to understand the complex relations of claims in the articles. ClaimDiff can be applied to provide a granular-level comparison between two articles. Document-level ClaimDiff enables to provide information about which arguments of the first article strengthen or weaken the views of the other. We provide a live demo for document-level extension with RoBERTa-large model.[11] As an example, the demo result of the above topic is shown in Appendix G.

In order for the real-world document-level comparison scenario, we evaluate our baseline models on the test-doc dataset, which follows the real-world label distribution. Specifically, the test-doc dataset includes all non-overlapping sentence pairs, which were originally filtered out for the test dataset construction as described in Section 3.2. The results are shown in Table 4. Note that we do not provide human performance on the test-doc, as obtaining human annotation over the whole article is costly. Aligning with previous observations, RoBERTa (MNLI) and LoRA with the DeBERTa-XXL achieves the best AUROC on document-level ClaimDiff-S and ClaimDiff-W, respectively. However, unlike previous results, T0 achieves the second-best F1 on test-doc ClaimDiff-S. One possible reason in that finetuned models have a high proportion of false positives in the full document setting due to distribution shift, whereas the zero-shot model seems to be more robust to it.

Although the fine-grained comparison is helpful for understanding contentious issues, looking over the whole article pair is costly. Future work includes providing summarized statistics from fine-grained comparisons. For example, the ratio of strengthening / weakening pairs can represent how much the two articles oppose each other. We can further extend ClaimDiff to compare more than two articles with summarized results, and even compare between different presses.

7 Conclusion

This paper presents ClaimDiff , a new benchmark dataset of 2.9k annotation to compare claims in news articles on contentious issues. Unlike the previous fact verification, ClaimDiff focues on *comparing* the nuance between claim pairs from trusted sources, whether one claim *strengthens* or *weakens* the other. We experiment with pre-trained language models in finetuning, parameter-efficient finetuning, and zero-shot approaches. The results show a significant room for improvement with over 19% absolute gap between human and model performance. We further suggest document-level ClaimDiff as a real-world application and show its

[11] https://www.claimdiff.com

potential by presenting the baseline performance on the test-doc dataset that follows the real-world distribution. We hope this initial study could pave the way for providing an analysis tool for article readers to obtain an unbiased understanding of contentious issues.

Limitations

First, most articles are crawled from the US and UK presses. This means the crawled data is English-only and regionally biased, limiting the scope and the diversity of issues. Extending our work to other languages and more regionally-diverse presses will be helpful for reducing such bias in our dataset.

Second, we suspect that there will be a non-trivial annotation bias in our dataset. We are concerned with the fact that all of our in-house annotators share the same cultural background and similar personal interest (given that the annotators volunteered to partcipate in this turking task). Furthermore, given that ClaimDiff-W is aiming to catch the subtle differences in the nuances of these professional news articles, it is very challenging for different annotators to have a common view, especially compared to ClaimDiff-S (which also explains why ClaimDiff-W human performance is much lower than that of ClaimDiff-S).

Third, since ClaimDiff is a sentence-level comparison task, it currently does not give information about the surrounding context of each sentence. This means inter-sentence dependency such as coreference often cannot be resolved. One way to work around this is to give an access to the full articles for each claim pair, but we have refrained from it in this work for simplicity (though we believe it will be interesting to see if the performance can be improved with such access).

Fourth, the size of ClaimDiff is relatively small compared to other fact verficiation datasets. This is mainly because its annotation process is quite challenging and requires a substantial amount of time. Future work includes expanding the size of ClaimDiff when additional budget is available.

Acknowledgements

We thank Yongrae Jo, Joel Jang, Hyunji Lee, Hanseok Oh, Soyoung Yoon, Hwisang Jeon and Gangwoo Kim for the useful discussion and feedback on the paper. This work was partly supported by KAIST-NAVER Hypercreative AI Center (80%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub, 20%).

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of EMNLP*.

Tariq Alhindi, Amal Alabdulkarim, Ali Alshehri, Muhammad Abdul-Mageed, and Preslav Nakov. 2021. AraStance: A multi-country and multi-domain dataset of Arabic stance detection for fact checking. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating stance detection and fact checking in a unified corpus. In *NAACL-HLT*.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *EACL*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *NAACL-HLT*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina

Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *NAACL-HLT*.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *SemEval-2017*, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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 *ACL*.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *NAACL-HLT*.

Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *WWW*.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *PCoNLL*, Hong Kong, China.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1).

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *ICLR*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *ICLR*.

Ye Jiang, Xingyi Song, Carolina Scarton, Ahmet Aker, and Kalina Bontcheva. 2021. Categorising fine-to-coarse grained misinformation: An empirical study of covid-19 infodemic. *arXiv preprint arXiv:2106.11702*.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *COLING*.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-ex: Improving sequence-to-sequence models for extractive rationale generation. In *EMNLP*.

Markus Leippold and Thomas Diggelmann. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of ACL-IJCNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *SemEval-2016*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *CIKM*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*.

Jeppe Nørregaard and Leon Derczynski. 2021. DanFEVER: claim verification dataset for Danish. In *NoDaLiDa*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. Zero: Memory optimizations toward training trillion parameter models.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *ACL-IJCNLP*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *EMNLP*.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Maxwell Weinzierl and Sanda Harabagiu. 2022. Identifying the adoption or rejection of misinformation targeting covid-19 vaccines in twitter discourse. In *WWW*.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *CICLing*, CICLing'05, Berlin, Heidelberg. Springer-Verlag.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*.

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*.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *EMNLP*.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3).

A Data Construction Process

This section describes the preprocessing step and the data construction process by human annotators. We construct ClaimDiff by 2 steps: (i) filtering non-overlapping claim pairs and (ii) annotating the pairs.

Removing Identifying Information We first preprocess the collected articles to remove identifying information, such as reporters' contact information. We remove personal information in a two-step procedure. First, we use automatic ways—regular expressions and pre-trained language models—to classify whether a sentence contains personal information. Then, after the automatic removal step, we manually inspect each sentence again and remove the sentence if it contains personal information.

Data Filtering Process we use MTurk[12] for filtering large amount of non-overlapping claims. Each worker is asked to solve Human Intelligent Tasks (HITs), which consist of 6 multiple-choice questions. Each HITs is composed of 1 quiz question to manage workers and 5 claim pairs extracted from news articles. The interface for a single question is presented in Figure 3. The reward for a single HIT is $0.18. We collect the responses from 3 different workers for a single example. If more than two workers choose the 'overlap' or 'large overlap', the pair are then considered as the 'overlapping' pair. If more than two workers choose 'small or no overlap', then the pair is considered as 'non-overlapping'. We process the annotation step for only the 'overlapping' claim pairs.

Data Annotation Process For the second annotation step, we separately hire 15 in-house expert annotators. We held two training sessions for in-house experts; one for providing guidelines and the other for solving example tasks. Each expert should pass the final quiz (15 out of 16 questions) after training sessions to start the main tasks. The interface for annotation is shown in Figure 4. Annotators are asked to choose the directional relation of a given pair and select the rationale that supports the relation. In the data construction process, we provide additional context information for a better understanding of the sentence. We offer $0.25 for a single example.[13]

[12] https://www.mturk.com

[13] We provide at least $7.5 per hour even if annotators submit less than 30 responses.

Model	F1	Precision	Recall
ClaimDiff-S			
RoBERTa-base	81.87	76.55	87.99
RoBERTa-large	84.70	76.17	95.38
RoBERTa (FEVER)	80.38	68.58	97.08
RoBERTa (MNLI)	83.12	78.26	88.62
LoRA (RoBERTa)	82.91	77.19	89.54
LoRA (DeBERTa)	82.60	75.68	90.91
ClaimDiff-W			
RoBERTa-base	43.24	34.78	57.14
RoBERTa-large	38.37	25.38	78.57
RoBERTa (FEVER)	18.29	10.39	76.19
RoBERTa (MNLI)	59.02	50.70	70.59
LoRA (RoBERTa)	43.18	41.30	45.24
LoRA (DeBERTa)	56.18	53.19	59.52

Table 5: Validation performance of finetuning and parameter-efficient finetuning baselines.

Hyperarameter	Search space
Learning rate	{5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6}
Warmup steps	{0, 50, 100, 150, 200}
Weight decay	{off, 1e-5}

Table 6: Search space for hyperparameters of finetuned RoBERTa.

B Implementation Details

For all experiments, we use PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020). Most of the experiments are conducted on 8 V100 GPUs except T5 for rationale extraction. The validation performance of each model is presented in Table 5.

B.1 RoBERTa (MNLI) / RoBERTa (FEVER)

RoBERTa (MNLI) and RoBERTa (FEVER) are finetuned RoBERTa-large models finetuned on MNLI and FEVER, respectively. We load the trained checkpoints for the MNLI model,[14] while me manually finetune RoBERTa-large[15] for FEVER. Following Nie et al. (2019), we convert FEVER into an NLI-style task, predicting only the labels among the given query and context. We train RoBERTa-large on NLI-style FEVER during 10 epochs with batch size of 32 and 100 warmup steps. The model is optimized by Adam(Kingma and Ba, 2015) optimizer with learning rate of $5e^{-5}$.

B.2 Relation Classification

Fientuned RoBERTa For finetuning experiments, we train models during 10 epochs with a batch size of 32 and 200 warmup steps. We find the best hyperparameters for each model using the

[14] https://huggingface.co/roberta-large-mnli

[15] https://huggingface.co/roberta-large

results of 3-fold cross-validation. We choose the best-working checkpoints and thresholds based on the validation F1 score. The search space for each hyperparameter is presented in Table 6. We use Adam optimizer for training. For ClaimDiff-S, we use learning rates of $5e^{-5}$ for RoBERTa-base and $5e^{-6}$ for others. ClaimDiff-W models are trained with learning rate of $1e^{-5}$ for RoBERTa (MNLI), and $5e^{-6}$ for other models.

LoRA We train RoBERTa-large and DeBERTa-XXL[16] with LoRA during 20 epochs with a batch size of 32, the learning rate of $1e^{-4}$, and 0.01 weight decay. We use the LoRA implementation released by the authors[17]. We use a linear scheduler for the learning rate schedule with a 0.1 warmup ratio. We set the rank as 8 with LoRA α of 16 for RoBERTa-base. For DeBERTa-XXL, rank and α are set to be 16 and 32.

T0 As the proposed tasks are binary classification tasks, T0 takes the input prompts and generates answers between ('yes', 'no') as prediction labels. We use pre-trained weights of T0[18] for zero-shot prediction. The input prompts for ClaimDiff-S and ClaimDiff-W are as follow:
(i) ClaimDiff-S

```
Claim A: {claim_a} \n\n
Claim B: {claim_b} \n\n
Does Claim A support Claim B? yes or no?
```

(ii) ClaimDiff-W

```
Claim A: {claim_a} \n\n
Claim B: {claim_b} \n\n
Does Claim A weaken Claim B? yes or no?
```

GPT-3 Similar to T0, we consider predictions of GPT-3 as correct if GPT-3 generates 'Yes' (when the label is 1) or 'No' (when the label is 0) in the outputs. We use *text-davinci-003* of the GPT-3 family in this work. The input prompts for ClaimDiff-S and ClaimDiff-W are as follow:
(i) ClaimDiff-S

```
Does A support B?: \n\n
A: {claim_a} \n\n B: {claim_b} \n\n
```

(ii) ClaimDiff-W

```
Does A weaken B?: \n\n
A: {claim_a} \n\n B: {claim_b} \n\n
```

[16] https://huggingface.co/microsoft/deberta-v2-xxlarge
[17] https://github.com/microsoft/LoRA
[18] https://huggingface.co/bigscience/T0

B.3 Rationale Extraction

MRC Models We train RoBERTa-base and RoBERTa-large with additional answer prediction layers during 3 epochs. Models are trained with Adam optimizer with a learning rate of $5e^{-5}$ and batch size of 32. We choose the best checkpoints based on validation TF1.

T5 We finetune T5-base[19] and T5-large[20] to directly generate a list of rationales. More formally, given claim pairs (c_1, c_2), we optimize models to obtain the list of rationale phrases. The model takes input as "explain claimdiff claim1: c_1 claim2: c_2", and is trained to generate the target sequence represented as "explanation: {rationale1} explanation: {rationale2} ···". For *class label* models, we additionally append "relation: r" to the input text as class information, where r is either one of *strengthen* or *weaken*.

We use T5 with a maximum input sequence length of 512 and a batch size of 8. All experiments are conducted on 4 Tesla M60 GPUs using ZeRO (Rajbhandari et al., 2019) stage-3 provided in DeepSpeed (Rasley et al., 2020) to reduce GPU memory usage. We train all models using the Adam optimizer with a constant learning rate of $1e^{-4}$. To obtain rationales, we perform beam search decoding using a beam size of 2.

C ClaimDiff Statisics

Table 7 shows top-15 presses and tags with their occurrence. Tags in ClaimDiff have long-tailed distribution, indicating ClaimDiff do not concentrate of specific topic.

D ClaimDiff Examples

The examples positive and negative examples of ClaimDiff are presented in Table 8. Note that the same pair can have different labels for ClaimDiff-S and ClaimDiff-W.

E Dataset Analysis

Figure 5 provides more detailed results of Section 3.4. For MNLI model, we gave the former claim as 'premise' and the later claim as 'hypothesis. For FEVER model predictions, the second claim are given as 'claim' and the former as 'evidence'.

[19] https://huggingface.co/t5-base
[20] https://huggingface.co/t5-large

F Error Analysis

We randomly sample 25 false negatives of RoBERTa-large predictions from ClaimDiff-S and ClaimDiff-W, respectively. Table 9 and Table 10 show each error category and its corresponding example in false negative errors.

G Live Demo Result

Figure 6 shows the screenshot of the demo and running results.

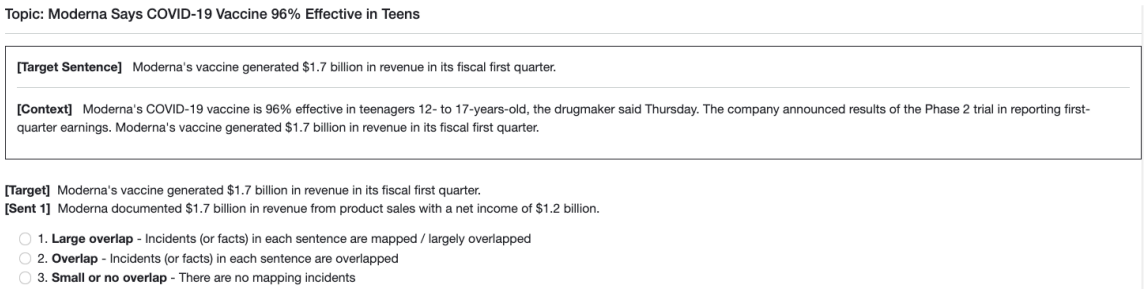


Figure 3: Interface for filtering task.

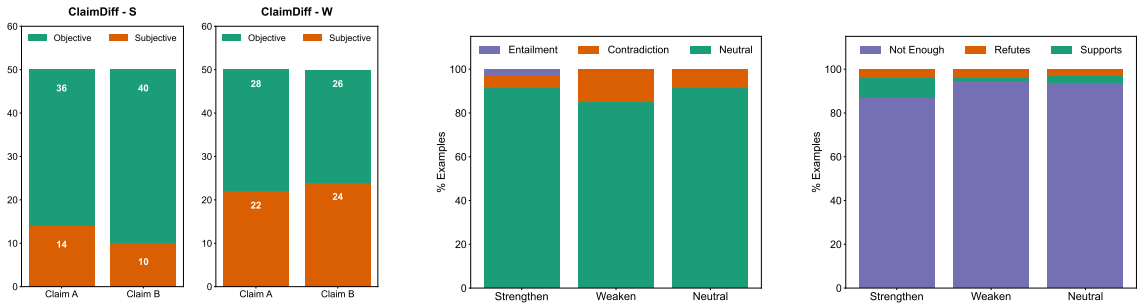
Figure 4: Screenshot of annotation task.

Press	Fox News (Online News) (32), CNN (Online News) (27), Washington Times (25), Politico (17), Associated Press (15), USA TODAY (12), New York Post (News) (11), NBC News (Online) (11), Reuters (11), The Hill (9), Vox (8), NPR (Online News) (8), The Guardian (7), National Review (6), New York Times (News) (6)
Tag	Elections (14), DonaldTrump (13), coronavirus (11), USSenate (9), Immigration (9), PresidentialElections (8), Technology (7), SupremeCourt (7), World (6), ViolenceinAmerica (5), MediaBias (5), WhiteHouse (5), Russia (5), MiddleEast (5), Business (5)

Table 7: Top-15 presses and tags included in ClaimDiff. The numbers indicate the occurrence of each press or tag.

Label	Example
	ClaimDiff-S
1	**Issue**: First-of-its-kind California Program Offers Virus Aid to People in the Country Illegally **Claim A**: Legal complaints lodged to try to stop the distribution of funds to illegal aliens were blocked, one by the California Supreme Court on May 6 and one by the Los Angeles Superior Court on May 5. **Claim B**: Applicants are eligible for the money if they demonstrate they are unauthorized, jobless as a result of the pandemic, and do not qualify unemployment programs or stimulus checks. **Rationale**: [Legal complaints lodged to try to stop the distribution of funds to illegal aliens were blocked,]
1	**Issue**: Pfizer Says its COVID-19 Vaccine is Safe, Effective for Kids Ages 5-11 **Claim A**: Coronavirus infections have risen "exponentially" among children across the United States, and now account for nearly 29% of all cases reported nationwide, the American Academy of Pediatrics reported last week. **Claim B**: "Since July, pediatric cases of COVID-19 have risen by about 240 percent in the U.S. - underscoring the public health need for vaccination," Pfizer's CEO Albert Bourla said in a statement. **Rationale**:[Coronavirus infections have risen "exponentially" among children across the United States,]
0	**Issue**: Hack Cuts Off Nearly 20% of US Meat Production **Claim A**: Any further impact on consumers will depend on how long JBS plants remain closed, analysts said. **Claim B**: The Colonial Pipeline, which provides 45% of the gas used in East Coast states, was hacked and temporarily shut down by East European hacker group DarkSide. **Rationale**: []
0	**Issue**: FDA Commissioner Acknowledges Misrepresenting Convalescent Plasma Data **Claim A**: The FDA made the decision based on data the Mayo Clinic collected from hospitals around the country that were using plasma on patients in wildly varying ways and there was no comparison group of untreated patients, meaning no conclusions can be drawn about overall survival. **Claim B**: Speaking at that press conference, Trump claimed that blood plasma treatment had cut COVID-19 mortality by 35%. **Rationale**: []
	ClaimDiff-W
1	**Issue**: FDA Commissioner Acknowledges Misrepresenting Convalescent Plasma Data **Claim A**: The FDA made the decision based on data the Mayo Clinic collected from hospitals around the country that were using plasma on patients in wildly varying ways and there was no comparison group of untreated patients, meaning no conclusions can be drawn about overall survival. **Claim B**: Speaking at that press conference, Trump claimed that blood plasma treatment had cut COVID-19 mortality by 35%. **Rationale**: [using plasma on patients in wildly varying, there was no comparison group of untreated patients, no conclusions can be drawn about overall survival.]
1	**Issue**: Facebook Changes Trending News **Claim A**: In a poll conducted by the media and data analysis site Morning Consult, only 48 percent of respondents said they had heard about the bias allegations against Facebook. **Claim B**: But the company also runs a "Trending Topics" section that promotes some stories, and that's where the bias charges focused. **Rationale**: [only 48 percent of respondents said they had heard about the bias allegations]
0	**Issue**: Hack Cuts Off Nearly 20% of US Meat Production **Claim A**: Any further impact on consumers will depend on how long JBS plants remain closed, analysts said. **Claim B**: The Colonial Pipeline, which provides 45% of the gas used in East Coast states, was hacked and temporarily shut down by East European hacker group DarkSide. **Rationale**: []
0	**Issue**: First-of-its-kind California Program Offers Virus Aid to People in the Country Illegally **Claim A**: Legal complaints lodged to try to stop the distribution of funds to illegal aliens were blocked, one by the California Supreme Court on May 6 and one by the Los Angeles Superior Court on May 5. **Claim B**: Applicants are eligible for the money if they demonstrate they are unauthorized, jobless as a result of the pandemic, and do not qualify unemployment programs or stimulus checks. **Rationale**: []

Table 8: Examples of ClaimDiff-S and ClaimDiff-W.



(a) Subjectivity Analysis (b) MNLI Model Predictions (c) FEVER Model Predictions

Figure 5: Analysis over claim pairs in ClaimDiff. (a) Subjectivity results for strengthening and weakening pairs with positive labels. (b) Prediction results of Roberta-large trained on MNLI dataset. (c) Prediction results of Roberta-large trained on FEVER.

ClaimDiff-S
(7 / 25) [Category 1] Entailment or Objective Example **Issue**: CDC Shortens School Distancing Guidelines to 3 Feet with Masks **Claim A**: Dr. Lawrence Kleinman, ..., said 3 feet is "probably safe" if schools are doing everything right - if everyone is wearing masks correctly at all times and washing their hands, and if ventilation is good. **Claim B**: In Utah , a study found that 86% of students wore masks in elementary school classrooms and very few passed the virus to others.
(4 / 25) [Category 2] Cuase-and-Effect **Issue**: Congress Estimates Social Security to Run Out by 2031 **Claim A**: Since no one is suggesting raising taxes to make up the lost revenue from Social Security, that additional $1 trillion would have more than doubled the fiscal year 2019 deficit. **Claim B**: But, ... removing the FICA tax as a funding source for Social Security, ... requires an increase in tax revenue of some fashion or another.
(5 / 25) [Category 3] Sharing the same nuances **Issue**: FDA Commissioner Acknowledges Misrepresenting Convalescent Plasma Data **Claim A**: Trump hailed the decision as a historic breakthrough even though the treatment's value has not been established. **Claim B**: The president also claimed that plasma "had an incredible rate of success" for treating COVID-19 patients, despite the fact that his own scientists and the FDA itself had expressed more reserved assessments.
(3 / 25) [Category 4] Lack of Context Information **Issue**: How the 4/20 Holiday is Celebrated During Coronavirus Pandemic **Claim A**: If you've been feeling anxiety over current events, it might be time to browse a few soothing CBD products - especially in honor of this week's cheeky 4/20 holiday, celebrated by cannabis lovers around the globe. **Claim B**: Part of that may have been in preparation for 4/20 celebrations.
(2 / 25) [Category 5] Supporting interview **Issue**: Jeff Sessions Hits Back At Trump **Claim A**: During a Thursday morning interview on "Fox & Friends," Mr. Trump renewed his criticism of Mr. Sessions, accusing him of allowing the Justice Department to undermine his administration. **Claim B**: "No, the truly unique thing here is that Sessions decided to actually speak up in his own defense."

Table 9: Cateogries and correspoding examples of false negatives in ClaimDiff-S. The number indicates how many examples fall into the category over 25 examples.

ClaimDiff-W
(15 / 25) [Category 1] Contradiction / Conflicts - (Type 1) Contradiction **Issue**: Election Systems Hacked by Russians **Claim A**: In this instance, the username and password information posted would only give individuals access to a localized, county version of the voting registration system, and not the entire state-wide system. **Claim B**: Hackers based in Russia were behind two recent attempts to breach state voter registration databases, fueling concerns the Russian government may be trying to interfere in the U.S. presidential election, U.S. intelligence officials tell NBC News.
(15 / 25) [Category 1] Contradiction / Conflicts - (Type 2) Conflicting Arguments **Issue**: FDA Commissioner Acknowledges Misrepresenting Convalescent Plasma Data **Claim A**: Though scientists and medical experts are in agreement that the emergency authorization would likely make it easier for certain hospitals and clinics to access plasma, a promising treatment strategy which uses antibodies of recovered patients, many expressed alarm Sunday over Trump's rhetoric. **Claim B**: Hahn had echoed Trump in saying that 35 more people out of 100 would survive the coronavirus if they were treated with the plasma.
(3 / 25) [Category 2] Lack of Context Information **Issue**: Negotiating the Fiscal Cliff **Claim A**: Obama expressed optimism as he took his case on the road here Friday, saying Democrats and Republicans "can and will work together." **Claim B**: The remarks came a day after the Obama administration unveiled details of a comprehensive package, widely rejected by Republicans, to avert the fiscal cliff.
(2 / 25) [Category 3] Opposing nuances **Issue**: CDC Issues Guidance for Fully Vaccinated Individuals **Claim A**: The guidance was "welcome news to a nation that is understandably tired of the pandemic and longs to safely resume normal activities," said Dr. Richard Besser, president and CEO of the Robert Wood Johnson Foundation and a former acting director of the CDC. **Claim B**: She stressed that everyone should continue to avoid nonessential trips, regardless of vaccination status.

Table 10: Cateogries and correspoding examples of false negatives in ClaimDiff-W. The number indicates how many examples fall into the category over 25 examples. Note that there are diverse patterns in contradiction / conflicts category, which makes ClaimDiff-W more challenging. We present two types contradiction / conflicts as examples.

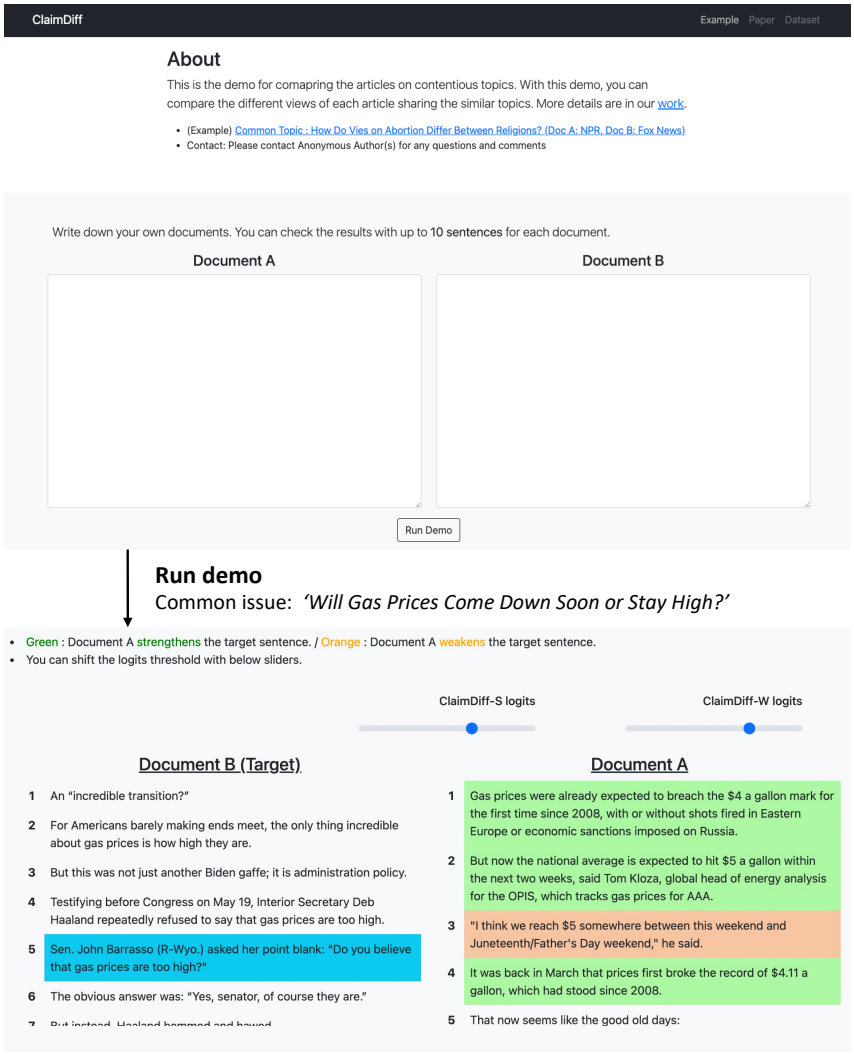


Figure 6: Demo results on the two articles about a topic, *'Will Gas Prices Come Down Soon or Stay High?'*. Sentences in green represent the claims that strengthen the 5-th sentence of document B. Sentence in orange indicates the claim that weakens the sentence of document B.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section name: Limitations

☑ A2. Did you discuss any potential risks of your work?
Section name: Limitations

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1 (Abstract), Section (Introduction)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4, Section 5, Appendix B

☑ B1. Did you cite the creators of artifacts you used?
Appendix B

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
[MIT license] Dataset: FEVER-NLI dataset Model checkpoints: RoBERTa-large / RoBERTa-base / RoBERTa-large-mnli / DeBERTa-XXL
[Apache 2.0] Model checkpoints: T0, t5-base, t5-large

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
All existing artifacts used in this paper are used in research purpose, which do not violate the intended use of CC BY-NC 4.0 license, MIT license, Apache 2.0 license.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix A

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Limitations

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4, Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4, Appendix B

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Not applicable. Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 3, Section 5, Appendix B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3, Appendix A

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Limitations