

# CMA-R: Causal Mediation Analysis for Explaining Rumour Detection

**Lin Tian**

RMIT University, Australia  
lin.tian2@student.rmit.edu.au

**Xiuzhen Zhang**

RMIT University, Australia  
xiuzhen.zhang@rmit.edu.au

**Jey Han Lau**

The University of Melbourne, Australia  
jeyhan.lau@gmail.com

## Abstract

We apply causal mediation analysis to explain the decision-making process of neural models for rumour detection on Twitter. Interventions at the input and network level reveal the causal impacts of tweets and words in the model output. We find that our approach CMA-R – Causal Mediation Analysis for Rumour detection – identifies salient tweets that explain model predictions and show strong agreement with human judgements for critical tweets determining the truthfulness of stories. CMA-R can further highlight causally impactful words in the salient tweets, providing another layer of interpretability and transparency into these blackbox rumour detection systems. Code is available at: <https://github.com/ltian678/cma-r>.

## 1 Introduction

There has been substantial work on understanding the inner workings of neural models via attention mechanisms (Clark et al., 2019), local surrogated approaches (Ribeiro et al., 2016; Lundberg and Lee, 2017; Kokalj et al., 2021) or integrated gradient based methods (Sundararajan et al., 2017). Existing works on explainable fake news or rumour detection by and large use attention weights to explain model decision (Shu et al., 2019; Khoo et al., 2020; Lu and Li, 2020; Li et al., 2021), but Pruthi et al. (2020) found that the use of attention as explanation is problematic: removing words with high attention appears to have little effect on the final prediction, suggesting that attention doesn't explain the decision process.

To address these limitations, in this paper, we propose CMA-R – Causal Mediation Analysis for Rumour detection – grounded in causal mediation analysis (CMA (Pearl, 2001), as illustrated in Figure 1) to interpret decisions for rumour detection models. CMA-R is a significant departure from existing interpretation methods, as it provides greater

explanatory power from assessing causal relations instead of correlations. Different from studies (Vig et al., 2020) that apply CMA to examine the causal structure from network components to predictions, we perform intervention in the input and network to determine the tweets and words that are *causally implicated* in the final prediction and verify them with human expert annotations. Using a rumour dataset that has been annotated by journalists to highlight critical tweets that determine the truthfulness of a story, we assess the salient tweets extracted by CMA-R and other interpretation methods (e.g. attention) and found that CMA-R yields better alignment with human judgements, empirically demonstrating that it is important to consider causality for explaining model decisions. CMA-R also allows us to highlight impactful words in those salient tweets, providing another mechanism to interpret rumour detection models.

The main contributions of this work are as follows:

- CMA-R is a novel application on interpreting rumour detection systems model decisions by performing interventions in the input and network that aims to identify tweets and words causally implicated in the final prediction.
- CMA-R can highlight impactful words in salient tweets via neuron level interventions, providing a refined mechanism for interpreting rumour detection models.
- Our findings show that CMA-R aligns more closely with human judgments on a journalist-annotated rumour dataset.

## 2 Related Work

We briefly summarise prior studies from three related areas: explainable artificial intelligence, causal mediation analysis and rumour detection.

Explainable artificial intelligence aims to create a suite of techniques to produce interpretable artificial intelligence systems, which are often driven by deep learning (Gunning et al., 2019). Broadly speaking there are two approaches: model-agnostic and model-specific methods. Model-agnostic approaches such as LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) and SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017; Kokalj et al., 2021) build local surrogate models to approximate the predictions of the original model. Model-specific techniques use feature visualisation (Vig, 2019) and attention mechanisms (Clark et al., 2019) to explain the decision-making process. Additionally, rationalisation-based approaches focus on generating textual explanations that rationalise a model’s decision. The explanations mimic human reasoning and provide narrative or rationale for why a model made a certain decision (Rajani et al., 2019; Pan et al., 2022; Liu et al., 2022, 2023; Chrysostomou and Aletras, 2022). It is not a way to explain a model’s internal decision-making processes, but a method for rationalising the behaviour and justifying its predictions.

Causal mediation analysis (CMA) aims to uncover cause-and-effect relationships, and its application to understanding deep learning models is emerging (Vig et al., 2020; Feder et al., 2022; Qian et al., 2021). CMA-R goes beyond understanding the correlations between the input and output, but instead attempts to the causal structure for model decisions. In this paper, we employ CMA-R to understand how intervention at both the word and neuron levels affect the model’s predictions.

Deep learning is the dominant approach for automatic detection of online rumours and fake news (Shu et al., 2019; Khoo et al., 2020; Lu and Li, 2020; Li et al., 2021). Attention mechanisms have been widely used to explain model decisions (Shu et al., 2019; Khoo et al., 2020; Lu and Li, 2020), but there is emerging evidence showing that correlation does not always constitute explanation (Jain and Wallace, 2019; Serrano and Smith, 2019; Pruthi et al., 2020).

### 3 Preliminaries

Let  $X = \{x_0, x_1, x_2, \dots, x_n\}$  be a set of events, where an event  $x_i$  consists of either: (1) a source tweet and its comments (Figure 2); or (2) a story with a set of source tweets and their comments (Fig-

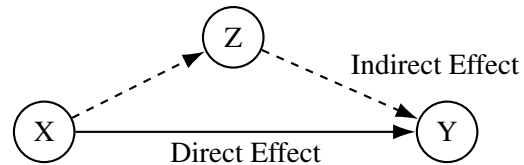


Figure 1: Casual mediation analysis.

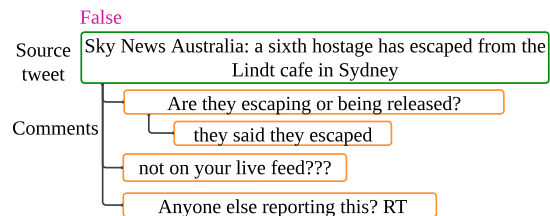


Figure 2: Labelled source tweet in PHEME.

ure 3). Each event  $x_i$  is associated with a rumour label  $y_i \in Y$ , where  $Y$  represents three rumour veracity classes (true, false or unverified). A rumour detection system is trained (with labelled data) to learn  $f : X \rightarrow Y$ .

## 4 Methodology

CMA-R allows us to analyse the change of a response variable ( $y$ ) following a treatment ( $x$ ) — e.g. in the biomedical domain this could mean the patient’s health outcome given a treatment — and it does so by considering *mediators* ( $z$ ), intermediate factors that produce an *indirect effect*. As shown in Figure 1, a mediator ( $z$ ) is added to take into account its indirect effect. Vig et al. (2020) introduce CMA as a means to explain the decision of a neural model, by viewing the model input as  $x$ , the model output (decision) as  $y$ , and the neurons in the model as  $z$ . In CMA-R,  $x$  represents an event and  $y$  a rumour label, and the tweets in  $x$  are encoded using a sequence network (e.g. BERT (Devlin et al., 2018)). The tweets in  $x$  may be concatenated as a string or represented as a graphs (to model the conversation structure), depending on the rumour detection model (Section 5.2).

### 4.1 Total Effects

To measure the causal impact of a tweet (or a set of tweets) in an event ( $x$ ) that contribute to a model prediction ( $y$ ), we can perform intervention by masking it out and computing the total effect:

$$TE = D(\mathbf{y}_{\text{null}}(x), \mathbf{y}_{\text{mask-text}}(x)) \quad (1)$$

where “null” and “mask-text” denote the intervention operations: the former performs no interven-

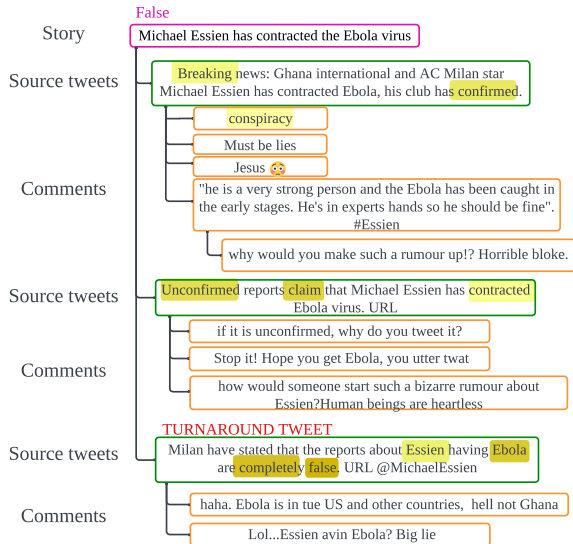


Figure 3: A labelled story in PHEME. Additional stories can be found in Appendix C.

tion and the latter masks out tweet(s) in the input (Figure 4 left);  $y$  represents the output probability distribution over the three veracity classes and  $D$  is a distance metric between two probability distributions (Section 4.3).

## 4.2 Indirect Effects

CMA-R also allows us to measure the causal impact of a neuron (or a set of neurons) by computing the indirect effect. The idea is to replace the value of a neuron in the pre-intervention network using that of the post-intervention network and measure how much that changes prediction. Formally:

$$IE = D(\mathbf{y}_{\text{null}}(x), \mathbf{y}_{\text{replace-neuron}}(x)) \quad (2)$$

where “replace-neuron” is the intervention operation for neuron replacement (Figure 4 right). Given that we use sequence networks (e.g. recurrent or transformer) to encode text, we can target neurons associated with words to measure the causal impact of each word, e.g. for a transformer encoder we can perform this replacement for neurons at different transformer layers that correspond to a word.

## 4.3 Distance Metric

Vig et al. (2020) use CMA for a task which has a binary outcome, and they propose computing the ratio between the probabilities of the positive class pre- and post-intervention to compute total/indirect effect. In our case (CMA-R), as we are dealing with a multi-class classification problem (3 veracity classes), we experiment with the following two

distance metrics for two probability distributions (Dwork et al., 2012):

$$T_1 = \frac{1}{2} \sum_{y \in Y} |y_{\text{null}}(x) - y_{\text{intervention}}(x)|$$

$$T_2 = e^{\max_{y \in Y} \log(\max(r_y, 1/r_y))}$$

where  $y_{\text{null}}(x)$  and  $y_{\text{intervention}}(x)$  denote the output probability of a label without and with intervention respectively and  $r_y = \frac{y_{\text{null}}(x)}{y_{\text{intervention}}(x)}$ . To rank the causal impact of tweets (total effect), we compute two rankings using the two distance metrics and sum the rankings to produce the final ranking. We rank the causal impacts of words (indirect effect) in the same way (i.e. via sum rank).

## 5 Experiment

### 5.1 Datasets

We use two variants of PHEME that contain veracity labels at two different levels: (1) source tweet (Figure 2; Kochkina et al. (2018));<sup>1</sup> and (2) story (Figure 3; Zubiaga et al. (2016)).<sup>2</sup> The former contains 29,387 labelled source tweets (with comments) while the latter has 46 labelled stories (each story can be interpreted as a news event that is linked to a number of related source tweets).<sup>3</sup> Each labelled story however, is also annotated with a “turnaround tweet” – the source tweet judged (by journalists) to be the critical tweet that determined the final veracity of a story.<sup>4</sup> We use the (larger) first PHEME variant to train a rumour classifier, and then apply the trained classifier to the (smaller) second PHEME variant to classify the stories and assess whether the salient source tweets extracted by CMA-R correspond to the ground truth turnaround tweets. Note that there is no overlap in terms of source tweets between the first and second PHEME variant, and so the rumour classifier has not “seen” any of the stories.

### 5.2 Models and Training Strategies

We experiment with three models with different architecture for encoding the tweets in  $x$ : (1) **one-tier transformer** uses RoBERTa (Liu et al., 2019) to

<sup>1</sup>figshare.com/articles/dataset/PHEME\_dataset\_for\_Rumour\_Detection\_and\_Veracity\_Classification/6392078

<sup>2</sup>figshare.com/articles/dataset/PHEME\_rumour\_scheme\_dataset\_journalism\_use\_case/2068650

<sup>3</sup>The description of a story, e.g. *Michael Esseien has contracted the Ebola virus* in Figure 3 is written by journalists.

<sup>4</sup>Technically, original dataset has 240 labelled stories, but only 46 of them has a turnaround tweet.

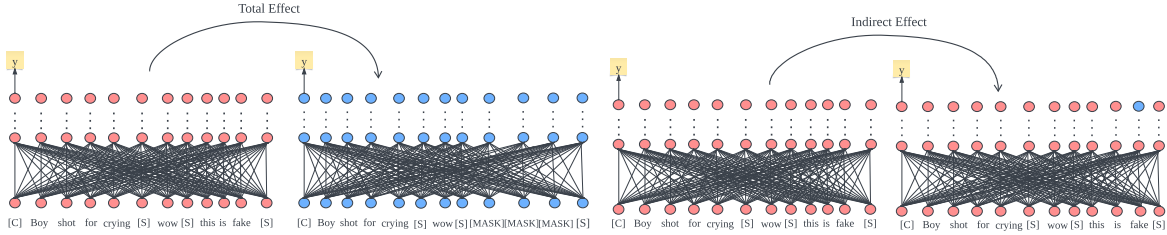


Figure 4: Total effect and indirect effect in CMA-R. [C] ([CLS]) and [S] ([SEP]) represent special tokens.

encode the tweets concatenated as a string; (2) **two-tier transformer** (Tian et al., 2022) uses BERT (Devlin et al., 2018) to encode each tweet separately and then another (randomly initialised) transformer to encode the sequence of [CLS] output embeddings from BERT; and (3) **DUCK** (Tian et al., 2022) uses BERT to encode each pair of parent-child<sup>5</sup> tweet and a graph attention network to encode the output from BERT to capture the conversation structure.<sup>6</sup> DUCK represents the current state-of-the-art for rumour detection.

In terms of training strategy, we explore two methods: (1) fine-tune using PHEME; and (2) fine-tune using Twitter15/16 and PHEME (in sequence). As Twitter15/16 is a larger labelled rumour dataset, we suspect the additional training would improve the models’ veracity prediction performance.

### 5.3 Baseline Interpretation Models

We test CMA-R with three other common baselines to extract salient tweets: (1) **attention**: we aggregate the attention weights for each word (one/two-tier transformer) or node (DUCK) and then rank each source tweet+comments by computing the average attention weight over their words (one/two-tier transformer) or nodes (DUCK); (2) **local**: we use LIME (Ribeiro et al., 2016) to compute word weights, and aggregate word weights in the same way as described before;<sup>7</sup>; (3) **gradient**: we compute word weights based on their gradients (Sundararajan et al., 2017) and aggregate word weights.

We further compare with three baseline systems for explainable fake news and rumour detection: (1) dEFEND (Shu et al., 2019) generates attention scores for both source tweets and their comments.

<sup>5</sup>Child tweet here means a replying comment.

<sup>6</sup>In the original paper the best DUCK variant is an ensemble that combines all three architectures.

<sup>7</sup>We use the following code for one/two-tier transformer and DUCK respectively: <https://github.com/cdpierce/transformers-interpret>, <https://github.com/mims-harvard/GraphXAI>.

The comment receiving the highest attention score is selected as the “turnaround tweet” – the key tweet that provides the most explanatory power in the context of a rumour. (2) GCAN (Lu and Li, 2020) does not explicitly identify the most explainable tweet in its original formulation. Attention scores are generated through its post and propagation attention mechanism. We adapted this by selecting tweets with the highest attention scores in this mechanism, assuming these to be the most relevant for explanation purposes. (3) StA-HiTPLAN (Khoo et al., 2020) provides post-level explanations based on the attention scores of the last layer. We used these post-level explanations to match back to the human-identified decision points in our datasets, assuming that higher attention scores correlate with greater explanatory relevance. All three baselines belong to attention-based approaches.

Model	F1	Turnaround Accuracy				
		R	A	L	G	C
Fine-tune with PHEME						
One-Tier	0.70	0.05	0.26	0.20	0.33	0.41*
Two-Tier	0.73	0.05	0.28	0.28	0.41	0.54*
DUCK	0.81	0.05	0.26	0.26	0.46	0.65*
dEFEND (Shu et al., 2019)	0.62	-	0.20	-	-	-
GCAN (Lu and Li, 2020)	0.72	-	0.28	-	-	-
StA-HiTPLAN (Khoo et al., 2020)	0.39	-	0.09	-	-	-
Fine-tune with Twitter15/16 and PHEME						
One-tier	0.72	0.05	0.26	0.20	0.37	0.43*
Two-tier	0.75	0.05	0.30	0.28	0.43	0.61*
DUCK	0.85	0.05	0.30	0.28	0.48	0.70*
dEFEND (Shu et al., 2019)	0.66	-	0.22	-	-	-
GCAN (Lu and Li, 2020)	0.75	-	0.28	-	-	-
StA-HiTPLAN (Khoo et al., 2020)	0.42	-	0.09	-	-	-

Table 1: Turnaround accuracy results. F1 denotes rumour classification performance. R: random baseline; A: attention; L: local; G: gradient; and C: CMA-R. An asterisk (\*) indicates that the result is statistically significant with  $p \ll 0.05$ . Detailed scores are in Appendix E.

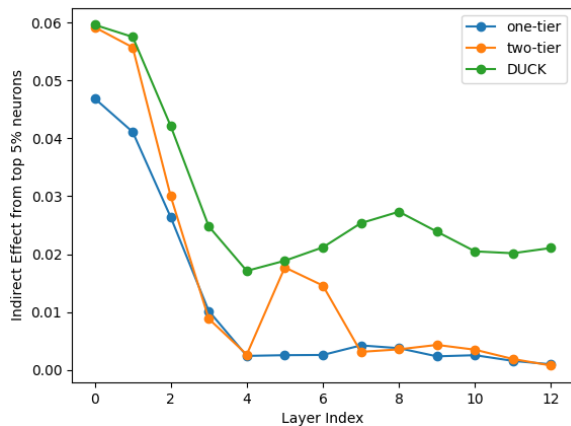


Figure 5: Indirect effects over different layers

## 6 Results

### 6.1 Turnaround Accuracy

We now assess how well the different interpretation methods pick up the correct turnaround tweets. Note that for CMA-R, when performing the “mask-text” intervention (Section 4.1) we mask each source tweet (and their associated comments) one at a time in order to determine which source tweet has the most causal impact. Table 1 presents the results. “R” denotes a random baseline where a random source tweet is chosen; 0.05 indicates on average 20 source tweets in a story. It is therefore a non-trivial task to identify the turnaround tweet.

We first look at the two fine-tuning strategies, and we see (without surprise) that the use of additional training data (Twitter15/16) improves rumour detection performance for all models, and that in turn leads to higher turnaround accuracy. Comparing the three models, DUCK is the clear winner here. Looking at the different interpretability methods (attention, local, gradient and CMA-R), we have a consistent observation: CMA-R is much more accurate at extracting the correct turnaround tweets, followed by gradient. Compared with existing explainable rumour detection approaches (Shu et al., 2019; Lu and Li, 2020; Khoo et al., 2020), we still can see that CMA-R better aligns with the human decision points. At a higher level, these results imply that it is important that we consider causal relations rather than correlations when interpreting model decisions.<sup>8</sup> We next present additional anal-

<sup>8</sup>In Appendix B, we provide further analyses where we consider only stories where a model have predicted the rumour veracity correctly (true or false). The general finding is broadly the same, where DUCK+CMA-R is the best combination in terms of veracity and turnaround prediction.

yses, and in these experiments we use Twitter15/16 and PHEME fine-tuned DUCK.

### 6.2 Salient Words

We use CMA-R to extract the most salient words by computing the indirect effects. When performing the “replace-neuron” intervention (Section 4.2), we replace the neurons for one transformer layer at a time, word by word. As such, we have a ranking of words for each layer, and we sum the rankings from the word embeddings and first six transformer layers. We highlight (in yellow) the most impactful words for a story in Figure 3. Interestingly, CMA-R extracts a number of intuitively critical words in the turnaround tweet, suggesting that it is focusing on the right words when making its decision.

### 6.3 Sparsity and Layer effects distribution

Following Vig et al. (2020) we also compute the indirect effects of the top neurons in different layers; results in Figure 5. In terms of the magnitude of indirect effects, DUCK seem to produce substantially higher effects. Across the layers, the earlier layers appear to have a much larger impact (this isn’t a surprising finding, as they are connected to more neurons in the network). Interestingly, though, we see a small bump in the middle layers of DUCK and two-tier transformer, which Vig et al. (2020) also found. In Appendix A, we present further analyses on the total effects.

## 7 Conclusion

We employed causal mediation analysis to understand the inner workings of rumour detection models. By performing interventions at the input and network levels, we show that our approach CMA-R can find tweets and words having the most causal impact for model decisions. To evaluate the “quality” of these insights, we train rumour detection models of differing complexity and compare CMA-R to current interpretation methods to assess how well the extracted salient tweets align with human judgements. Empirical results demonstrate that CMA-R is consistently the best method, suggesting that causal relations, rather than correlations, can better interpret model decisions. CMA-R provides further mechanism to hone in on the words for the most causal impact, and qualitative analysis reveals that the best rumour detection model is focusing on intuitively important words when determining the veracity of a story.

## 8 Limitations

We acknowledge that the size of our test data (story-annotated PHEME) is relatively small (46 instances), and this points to the laborious and difficult nature of the annotation task. That said, we contend that our results constitute one of the first studies in rumour detection that attempts to empirically validate the quality of insights produced by interpretation methods. To ensure the robustness of our results, we have conducted significance tests (results included in Appendix E).

While our work primarily focuses on applying causal mediation analysis to text-based rumour detection models, it is important to acknowledge that we did not apply user-based or propagation-based interventions in this particular study. However, the emphasis on text-based analysis provides a foundation for future investigations that can extend our methodology to encompass other methods and incorporate a more comprehensive understanding of rumour detection systems.

## Acknowledgement

This research is supported in part by the Australian Research Council Discovery Project DP200101441. Lin Tian is supported by the RMIT University Vice-Chancellor PhD Scholarship (VCPS).

## References

- George Chrysostomou and Nikolaos Aletras. 2022. Flexible instance-specific rationalization of nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10545–10553.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert’s attention. *ACL 2019*, page 276.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.
- Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.
- David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. Xai—explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.
- Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.
- Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8783–8790.
- Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413.
- Enja Kokalj, Blaž Škrlj, Nada Lavrač, Senja Pollak, and Marko Robnik-Šikonja. 2021. Bert meets shapley: Extending shap explanations to transformer-based classifiers. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21.
- Jiawen Li, Shiwen Ni, and Hung-Yu Kao. 2021. Meet the truth: Leverage objective facts and subjective views for interpretable rumor detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 705–715.
- Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Xinyang Li, YuanKai Zhang, and Yang Qiu. 2023. MGR: Multi-generator based rationalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12771–12787, Toronto, Canada. Association for Computational Linguistics.
- Wei Liu, Haozhao Wang, Jun Wang, Ruixuan Li, Chao Yue, and YuanKai Zhang. 2022. Fr: Folded rationalization with a unified encoder. *Advances in Neural Information Processing Systems*, 35:6954–6966.
- 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*.
- 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 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

- Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Sicheng Pan, Dongsheng Li, Hansu Gu, Tun Lu, Xufang Luo, and Ning Gu. 2022. Accurate and explainable recommendation via review rationalization. In *Proceedings of the ACM Web Conference 2022*, pages 3092–3101.
- Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the seventeenth Conference on Uncertainty in Artificial Intelligence (UAI'01)*, pages 411–420.
- Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793.
- Shangshu Qian, Viet Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. 2021. Are my deep learning systems fair? an empirical study of fixed-seed training. *Advances in Neural Information Processing Systems*, 34:30211–30227.
- Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.
- Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.
- Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.
- Lin Tian, Xiuzhen Jenny Zhang, and Jey Han Lau. 2022. Duck: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949.
- Jesse Vig. 2019. Bertviz: A tool for visualizing multi-head self-attention in the bert model. In *ICLR workshop: Debugging machine learning models*.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.
- 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):e0150989.

## A Magnitude of Total Effects

Model	Params	$T_1$	$T_2$
One-tier	125M	0.27	0.12
Two-tier	165M	0.30	0.19
DUCK	143M	0.73	0.55

Table 2: Average Total Effects.

To calculate the total effect for each model, we compute the average total effects by aggregating the individual effects across all 46 test instances. These effects represent the cumulative influence of the model neurons on the interventions. Table 2 shows the magnitude of average total effects (over source tweets and stories) for the two distance metrics. Interestingly, we find that the total effects using DUCK appears to be substantially larger.

## B Turnaround Accuracy

To better understand the effectiveness of causal mediation analysis as a way to explain model decisions, we further measure its performance under the conditional scenario. In this case, we do care about whether the model correctly predicted the rumour’s truthfulness. Since resolving tweets lead to a rumour being labelled as true or false, we can measure how accurately the model predicts this. In this scenario, we look at both how well the model predicts the rumour’s truthfulness and how accurately it identifies the key turning points in the conversation. The results are shown in Table 3.

## C Labelled Samples in PHEME

In order to provide a better understanding of the dataset utilised in our experiments, this section will

Model	F1	Conditional TRUE (27)					Conditional FALSE (19)				
		#TP	Attention	Local	IG	CMA-R	#TP	Attention	Local	IG	CMA-R
Fine-tune with PHEME											
One-Tier	0.70	17	0.18	0.24	0.24	0.24	11	0.09	0	0.36	0.64
Two-Tier	0.73	18	0.22	0.22	0.28	0.33	12	0.08	0.17	0.42	0.58
DUCK	0.81	23	0.26	0.35	0.43	0.52	14	0.14	0.29	0.50	0.57
Fine-tune with Twitter15/16 and PHEME											
One-tier	0.72	20	0.20	0.10	0.20	0.30	12	0.17	0.08	0.33	0.58
Two-tier	0.75	21	0.19	0.29	0.48	0.57	13	0.08	0.23	0.46	0.62
DUCK	0.85	23	0.35	0.35	0.52	0.61	15	0.13	0.27	0.47	0.60

Table 3: Turnaround accuracy results. F1 denotes rumour classification performance. #TP represents the number of correct classified instances.

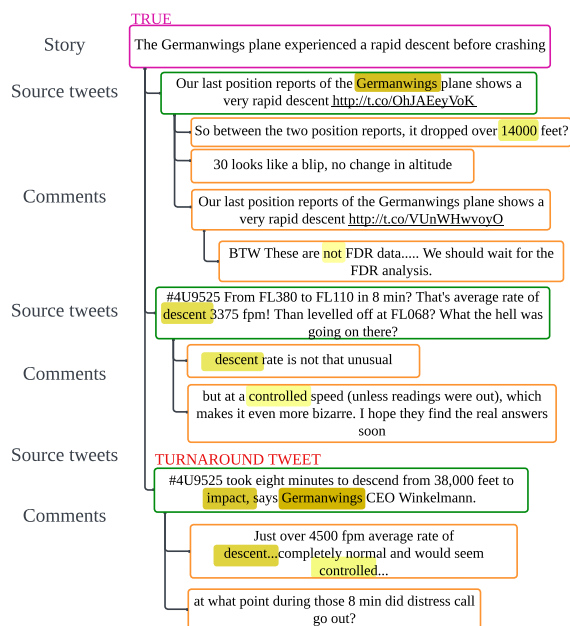


Figure 6: A labelled true story in PHEME.

further include labelled story samples (Figure 6 and Figure 7), supplementing the example presented in Figure 3 of the main manuscript, ensuring consistency of our findings.

## D Hyper-parameter Details

To fine-tune the base rumour detection model, we use the development set of the dataset for tuning hyper-parameters for each model. The detailed searched hyper-parameters are listed in Table 4.

## E Statistical Test

In the qualitative analysis, we conducted significance tests to validate the performance improvements across three types of interpretability mod-

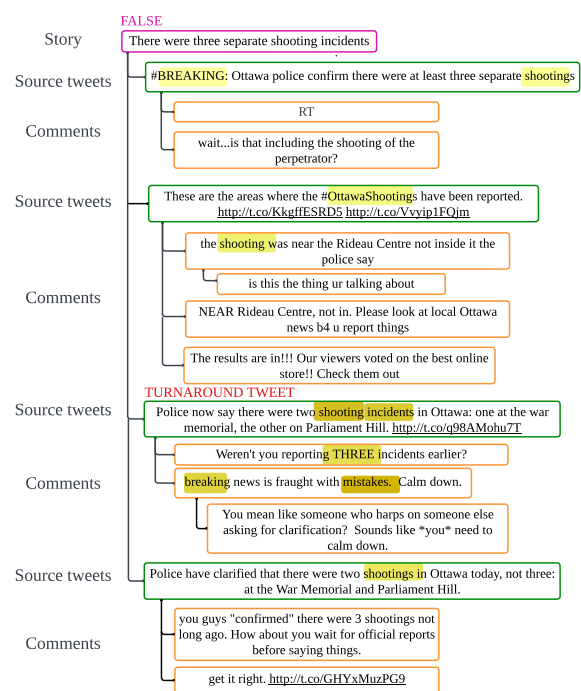


Figure 7: A labelled false story in PHEME.

els. We conducted Man-Whitney tests on accuracy for identifying turnaround posts. Results show that CMA-R is statistically significantly better than other interpretability models  $p - value \ll 0.05$ . Results are shown in Table 6.



Model	Base Encoder	Learning Rate	Dropout Rate
One-tier Transformer	RoBERTa	[3e-5, 5e-5]	[0.4-0.5]
Two-tier Transformer	BERT	[2e-5,5e-5]	[0.5-0.6]
DUCK	BERT	[1e-5, 5e-5]	[0.1-0.2]

Table 4: Hyper-parameters.

Dataset	# source tweet	#comments	# stories
PHEME (Kochkina et al., 2018)	6,245	98,929	–
PHEME (Zubiaga et al., 2016)	7,507	32,154	240

Table 5: Datasets Statistics.

Model	Pairs	P-value
One-Tier	CMA-R vs Random	0.00016
One-Tier	CMA-R vs Attention	0.00348
One-Tier	CMA-R vs Local	0.00138
One-Tier	CMA-R vs Gradient	0.02925
Two-Tier	CMA-R vs Random	0.00015
Two-Tier	CMA-R vs Attention	0.00040
Two-Tier	CMA-R vs Local	0.00055
Two-Tier	CMA-R vs Gradient	0.01040
DUCK	CMA-R vs Random	0.00016
DUCK	CMA-R vs Attention	0.00040
DUCK	CMA-R vs Local	0.00040
DUCK	CMA-R vs Gradient	0.00467

Table 6: Mann-Whitney U test results.