

Can We Identify Stance Without Target Arguments? A Study for Rumour Stance Classification

Yue Li, Carolina Scarton

University of Sheffield
211 Portobello, Sheffield, UK
yli381, c.scarton@sheffield.ac.uk

Abstract

Considering a conversation thread, rumour stance classification aims to identify the opinion (e.g. agree or disagree) of replies towards a *target* (rumour story). Although the target is expected to be an essential component in traditional stance classification, we show that rumour stance classification datasets contain a considerable amount of real-world data whose stance could be naturally inferred directly from the replies, contributing to the strong performance of the supervised models without awareness of the target. We find that current target-aware models underperform in cases where the context of the target is crucial. Finally, we propose a simple yet effective framework to enhance reasoning with the targets, achieving state-of-the-art performance on two benchmark datasets.

Keywords: rumour stance classification, rumour analysis on social media, stance classification

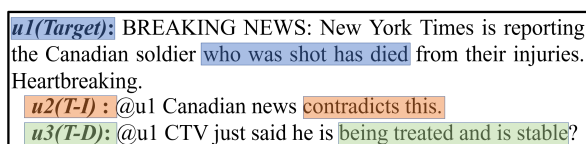
1. Introduction

Automatic stance classification that aims to identify the type of an expressed opinion towards a single or multiple *targets*, plays a key role in many Natural Language Processing (NLP) applications, such as rumour analysis (Zubiaga et al., 2016). A target could be a person, an organisation, or rumour story, depending on the use case (Hossain et al., 2020; Zubiaga et al., 2016; Ferreira and Vlachos, 2016; Allaway and McKeown, 2020). The target plays a fundamental role in stance classification, being expected to appear either explicitly or implicitly, making it a key difference from sentiment analysis that can be framed as target-independent (Küçük and Can, 2020; Liu et al., 2022).

Previous work shows that a BERT-based model, without awareness of the target, achieves comparable or even better performance than target-aware models on many stance classification datasets, due to spurious sentiment- and lexicon-stance correlations in the training sets (Kaushal et al., 2021). Similar results are observed in other context-dependent tasks, such as Natural Language Inference and Argument Reasoning Comprehension, where models without background knowledge achieve an impressive performance due to spurious or superficial cues in the datasets (Poliak et al., 2018; Niven and Kao, 2019).

In this paper, we further analyse the above phenomenon for *rumour stance classification* on Twitter. Given a conversation initialised by a rumourous *source tweet*, this task aims to classify the stance of each reply towards the rumour into *support*, *deny*, *query* and *comment*.¹ The vague-

¹The target of rumour stance classification is the rumour story by task definition, but these are not given in



u1(Target): BREAKING NEWS: New York Times is reporting the Canadian soldier who was shot has died from their injuries. Heartbreaking.
u2(T-I): @u1 Canadian news contradicts this.
u3(T-D): @u1 CTV just said he is being treated and is stable?

Figure 1: Example of Target-Independent (T-I) and Target-Dependent (T-D) direct replies that *deny* a target from Gorrell et al. (2019).

ness and lack of specificity in the reply tweets result in the disparity between rumour stance classification and traditional stance classification datasets. For instance, in Figure 1, one can reasonably deduce that the reply from *u2* disagrees with the target before reading the content of the target. This is in contrast to traditional stance classification where the stance may vary for different targets, making it always essential to consider them (e.g., Sobhani et al., 2017; Conforti et al., 2020).

We empirically show that the strong behaviour of models without awareness of the target (dubbed *target-oblivious*) could be explained by the existence of the reply posts whose stance can be naturally inferred without knowing the target.² More importantly, we demonstrate that current state-of-the-art target-aware models lack reasoning with the target, performing unexpectedly poorly on the cases when the target is necessary. Based on our observations, we propose a simple yet effective

the datasets. Instead, they are implied by the source tweets. In this work we use the terminology *target* to indicate the source tweet, because it is treated as the *target* in data annotation and applications (Zubiaga et al., 2016; Hardalov et al., 2022; Kaushal et al., 2021)

²Annotations can be found at: <https://github.com/YLi999/Target-Annotations-RumourEval>

tive framework which would benefit from the target-oblivious model and would also enhance the reasoning with the targets.

2. The Role of Target Arguments

We conduct an annotation study by categorising the replies into *target-dependent* (i.e. target is essential for stance inference) and *target-independent* (i.e. target is unnecessary for stance inference). We then evaluate various models trained with or without awareness of the target (i.e. *target-aware* and *target-oblivious* models).

2.1. Data Annotation

Dataset Three established English datasets are available for rumour stance classification on social media: *PHEME* (Zubiaga et al., 2016), *RumourEval 2017* (Derczynski et al., 2017) and *RumourEval 2019* (Gorrell et al., 2019). RumourEval 2017 consists of the English PHEME dataset, and RumourEval 2019 is an extension of the 2017 dataset. Therefore, we consider the largest RumourEval 2019 dataset.³ RumourEval 2019 training and validation sets consist of conversations regarding rumour stories which emerged during breaking news (e.g., Germanwings plane crash, and shooting in Ottawa), and the test data contains unseen rumours about natural disasters. The target of the stance, rumour story, is implied by the source tweet that initialises the conversation. Hence we consider the source tweet as the target. Among the four stances, *support* and *deny* classes are the most informative for rumour verification, while the *comment* class is the least useful (Scarton et al., 2020). Therefore, we annotate all the replies in *support*, *deny* and *query* classes in the validation and test sets, with 50 randomly sampled *comments* from each set.

Annotation Process Two annotators manually categorised each reply into either target-independent or -dependent, by answering one question: “do you think you need the source tweet to infer the stance of this reply?” Aiming to validate the annotations, annotators were also asked to classify the stance of the tweets. We then compared their assigned class with the gold standard label and, if they differed, we altered their annotation from target-independent to -dependent. Annotators did not have access to the source tweet and the tweets from validation and test sets were shuffled before annotation. The inter-annotator agreement is of 72.5% and Cohen’s Kappa is 0.565.

³The dataset contains Twitter and Reddit. To alleviate the impact of text length, we focus on the Twitter data only

Dataset	Support	Deny	Query	Comment
Validation	20 (29%)	35 (51%)	42 (40%)	0 (0%)
Test	12 (13%)	66 (72%)	17 (30%)	0 (0%)

Table 1: Number of target-independent tweets in each stance in the validation and test sets (proportion in brackets).

Result We observe a significant amount of data whose stance can be deduced without knowing the specific rumour story (Table 1), especially in the *deny* and *query* classes. More than 50% *denies* are target-independent in the validation and test sets. Target-independent *denies* are tweets that directly cast doubt with negation words (e.g. “Fake news”, “This is false”). The *queries* tend to be target-independent, since most of them are interrogative sentences asking for more evidence. However, the annotators did not identify many of them due to the ambiguity or non-informativeness of the texts (e.g., “blood clot?”, “WHAT?”). Most of the target-independent *supports* are retweets and quote tweets, whose context is self-contained. Tweets in the *comment* class are less relevant to the veracity of the rumour story, however, determining their relevance normally necessitates reasoning with the rumour story itself. We present more examples of target-independent and -dependent tweets in the Appendix A.

2.2. Model Evaluation

Given a source tweet (s_i), reply tweet to classify (r_i), other replies in the conversation (o_i) and stance label (l_i), we consider two types of supervised models: target-oblivious ($f(r_i) \rightarrow l_i$) and target-aware ($f(s_i, r_i)$ or $f(s_i, r_i, o_i) \rightarrow l_i$) models. We also evaluate a recent large language model (LLM) in zero-shot setting.

2.2.1. Experimental Setups

Target-oblivious Models We fine-tune different transformer-based models, whose input is the reply tweet ($f(r_i)$). We present the results using BERTweet (Nguyen et al., 2020) (experiments with BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) achieved similar performance).

Target-aware Models We fine-tune BERTweet, which takes as input both source and reply tweets ($f(s_i, r_i)$). We also evaluate four competitive systems that model the whole conversation thread ($f(s_i, r_i, o_i)$):⁴ (1) The winner of the RumourEval 2019 shared task, i.e. *BLCU-NLP* (Yang et al.,

⁴Performances of these models are lower than the figures reported in their original paper. The reason is that we do not consider the stance of the source tweet

Type	Model	Full set	Target-dependent subset			Target-independent subset		
		wF_2	wF_2	$F_2(S)$	$F_2(D)$	wF_2	$F_2(S)$	$F_2(D)$
Target-oblivious	BERTweet	0.477	0.346	0.294	0.206	0.749	0.615	0.894
Target-aware	BERTweet	0.435	0.329	0.313	0.167	0.635	0.464	0.778
	BLCU-NLP	0.371	0.223	0.080	0.217	0.399	0.000	0.737
	BUT-FIT	0.309	0.176	0.020	0.047	0.371	0.102	0.495
	Branch-LSTM	0.150	0.139	0.020	0.048	0.142	0.102	0.056
	Hierarchical-BERT	0.235	0.137	0.065	0.017	0.234	0.017	0.293
LLMs	LLaMA (reply)	0.256	0.227	0.390	0.000	0.319	0.417	0.093
	LLaMA (source & reply)	0.419	0.318	0.326	0.234	0.685	0.678	0.714

Table 2: Model performance over the full test set, target-dependent and -independent direct replies (averaged over experiments.). $F_2(S)$ and $F_2(D)$ denote the F_2 scores over *support* and *deny* classes, respectively. Highest performance is in bold, with statistical significance (t test, p value<0.05).

	Target-dependent subset				Target-independent subset			
	Support	Deny	Query	Comment	Support	Deny	Query	Comment
Mask Source Tweet	40.3	69.9	98.7	85.7	43.0	90.8	98.7	89.0
Shuffle Source Tweet	54.1	82.9	93.6	90.9	57.7	94.9	97.3	83.1

Table 3: The proportion (%) of target-aware BERTweet predictions of direct replies in each class that are not influenced by the masking or shuffling of the source tweets.

2019); (2) *BUT-FIT* (Fajcik et al., 2019), the second place in the 2019 shared task; (3) *Hierarchical-BERT* (Yu et al., 2020), achieving state-of-the-art performance (Hardalov et al., 2022) on the RumourEval 2017 dataset (Derczynski et al., 2017); (4) *Branch-LSTM* (Kochkina et al., 2017), the winner of the RumourEval 2017 shared task and the baseline model for the 2019 task.

LLMs We experiment with the OpenAssistant LLaMA-Based Model (Köpf et al., 2023).⁵ We compare the performance between two scenarios: (i) when the source tweet is provided (*LLaMA (source + reply)*) and (ii) when it is not (*LLaMA (reply)*).⁶

Evaluation We adopt the weighted F_2 score proposed by Scarton et al. (2020), which gives higher weights to the *support* and *deny* classes, being more adequate to rumour stance classification.

2.2.2. Results

As shown in Table 2, not surprisingly, all the models achieve better results on the target-independent samples, since they normally contain explicit stance-associated words or signals, especially for the *deny* and *query* classes. The target-oblivious model exhibits strong performance over target-independent tweets, indicating that its performance can be attributed to the existence of these samples in the dataset.

towards rumour, mainly belonging to the support class.

⁵<https://huggingface.co/OpenAssistant/oasst-sft-6-llama-30b-xor>

⁶Due to ethical considerations regarding the exposure of personal data (e.g., to ChatGPT), we opt to use an open-source LLM which was downloaded and hosted on our own server.

We expected that target-aware models, especially the ones that consider the whole conversation information, would perform significantly better than target-oblivious models on the target-dependent tweets for which the context of the source tweet is essential. However, among the target-dependent *supports* and *denies* that the target-oblivious BERTweet couldn't identify, BUT-FIT, Branch-LSTM and Hierarchical-BERT fail to correctly predict any of them as well, casting doubt on the usefulness of these target-aware approaches. BLCU-NLP is the only conversation-based system that outperforms the target-oblivious model over the target-dependent *denies*, likely due to their data augmentation for this class. But its performance over the target-dependent *supports* is rather disappointing.

Target-aware BERTweet shows strength on detecting target-dependent *supports*, when compared with its target-oblivious counterpart; however, it falls behind on the *deny* class. The existence of negation words (e.g., “not”) in the target-dependent *denies* may contribute to the good generalisation of target-oblivious BERTweet.

LLaMA exhibits competitive results, achieving best performance on the target-dependent samples in the *support* and *deny* classes. However, gaps still exist between the fine-tuned BERTweet models on the full test set. Without the source tweet, the performance drops significantly, except for the target-dependent *supports*.

2.2.3. Target perturbations

Aiming to further investigate the role of the target in target-aware models, we experiment with two perturbations during inference: (1) Masking: the entire source tweet is replaced by a white space; (2) Shuffling: the original source tweet is replaced

by a source tweet related to another rumour story so that the reply and “new” source tweets are mismatched. Both approaches should significantly change the model performance over the target-dependent tweets, provided the source tweet is properly reasoned with. We expect the *comment* class to be less impacted because the irrelevance between source and reply tweets should be considered as *comment*. We discuss the results of the target-aware BERTweet, since it is the best performing model in this category (other models showed similar results).

Masking or shuffling the source tweets has minimum impact over the predictions for the *deny*, *query* and *comment* classes (Table 3). More than 69% of predictions in each class stay the same, no matter whether the target is essential or not. For the *support* class in which target-aware BERTweet achieves better results over target-dependent samples, 40% to 60% of predictions do not change. The results suggest that target-aware models may be overfitting towards the replies, behaving like a target-oblivious model.

3. Ensemble-based Framework

Equipped with the observation of target-independent cases and the lack of reasoning with the target in target-aware models, we propose a simple yet effective ensemble-based framework to leverage the advantage of the target-oblivious model meanwhile improving the performance over the target-dependent samples.

We assume a pre-trained target-oblivious model ($f(r_i; \theta) = p_i$). The aim is to adopt an ensemble with a target-aware model ($f'(s_i, r_i; \theta') = q_i$) where p_i and q_i are posterior probability distribution over the four stance classes for a sample i with a pair of source (s_i) and reply (r_i) tweets. To encourage the target-aware model to learn from target-dependent samples during training, we propose a cross-attention based architecture with a sample re-weight mechanism.

Siamese Network with Cross-attention We utilise a siamese pre-trained transformer-based network (Reimers and Gurevych, 2019) to encode the source (s_i) and reply (r_i) tweets. Then, to explicitly indicate the importance of the tokens in the reply representation (h_{r_i}) with respect to the source representation (h_{s_i}), we calculate the cross-attention (Vaswani et al., 2017) between them, with h_{s_i} as the key and value, and h_{r_i} as the query.

Sample Re-weight We train the model on weighted data, where the weight of instance i is $1 - p_{y_i}$ (p_{y_i} is the posterior probability assigned to the true label y_i) (Clark et al., 2019). The intuition

is to encourage the target-aware model to focus on potential target-dependent examples that the target-oblivious model gets wrong.

Implementation Target-oblivious and -aware models are based on BERTweet but our method can be easily generalised to other pre-trained language models. The optimal target-oblivious model is chosen based on the validation set.

3.1. Experimental Setup

Datasets We validate our proposed framework on two benchmark datasets: RumourEval 2017 and 2019 datasets.

Comparing Baselines We compare with the Pretext Task-based Hierarchical Contrastive Learning model (PT-HCL) (Liang et al., 2022). To the best of our knowledge, PT-HCL is the only study that exploits “target-invariant/-specific features” (Liang et al., 2022) in traditional stance classification. We also present ablations for our proposed method, by removing the sample re-weighting mechanism (*w/o weight*), replacing the cross-attention by self-attention on the concatenation of the source and reply tweet representations (*w/o cross-att*), or both simultaneously (*w/o weight, cross-att*). Performance over RumourEval 2019 dataset is comparable with models in Table 4. As for RumourEval 2017, we also compare with its state-of-the-art model (Hierarchical-BERT), target-oblivious and -aware BERTweet and OpenAssistant LLaMa.

3.2. Results

As shown in Table 4, our proposed approach outperforms PT-HCL on both datasets, also surpassing other models. Removing sample weights or cross-attention would reduce the model performance, indicating their contribution.

Method	2019 dataset	2017 dataset
PT-HCL	0.452	0.431
Hierarchical-BERT	0.235	0.275
LLaMA	0.419	0.314
Target-oblivious BERTweet	0.477	0.425
Target-aware BERTweet	0.435	0.426
Proposed Method	0.510	0.452
w/o weight	0.458	0.421
w/o cross-att	0.438	0.417
w/o weight, cross-att	0.436	0.419

Table 4: Averaged wF_2 over experiments for two datasets. Highest performance is in bold, with statistical significance between the proposed method (t test, p value <0.05).

We also evaluate our proposed method and its ablations on target-dependent and -independent

	Target-dependent subset			Target-independent subset		
	wF_2	$F_2(S)$	$F_2(D)$	wF_2	$F_2(S)$	$F_2(D)$
Proposed Method	0.396	0.399	0.211	0.802	0.732	0.901
w/o weight	0.346	0.326	0.197	0.680	0.532	0.827
w/o cross-att	0.355	0.328	0.210	0.669	0.537	0.802
w/o weight,cross-att	0.314	0.322	0.191	0.627	0.390	0.804

Table 5: wF_2 scores of our proposed method and ablations over the target-dependent and -independent subsets of RumourEval 2019 test set. Highest performance is in bold, with statistical significance between "w/o weight,cross-att" (t test, p value<0.05).

subsets, as shown in Table 5. Comparing with Table 2, our model achieves the best results on both target-dependent and -independent examples, confirming that our proposed framework could not only benefit from the target-oblivious model but also enhance the inference between source and reply tweets. Furthermore, ensemble with either cross-attention or sample-weighting based target-aware model could improve the performance on average, if we compare the results of ablations. Sample-weights (w/o cross-att) could guide the target-aware model to focus more on the instances that target-oblivious model struggles with, resulting in more improvement over the target-dependent subsets than the method with only cross-attention (w/o weight). However, the differences are not statistically significant.

4. Conclusion

In this paper, we explore the role of the target in rumour stance classification. Our study suggests the strong performance of target-oblivious models could be explained by the existence of target-independent texts in real-world data. We point out the unexpected weakness of the target-aware models and consequently propose a cross-attention based architecture with a sample re-weight mechanism, achieving the best results on two benchmark datasets. We also release our annotations of target-dependent or -independent replies to facilitate future research and model evaluations. Finally, we argue that research in this area (and other textual entailment tasks) should conduct a thorough data analysis in order to fully understand models' performance, going beyond automatic metrics results.

5. Acknowledgements

This work is funded by the European Union under action number 2020-EU-IA-0282 and agreement number INEA/CEF/ICT/A2020/2381686 (EDMO Ireland).⁷ and by EMIF managed by the Calouste

⁷<https://edmohub.ie>

Gulbenkian Foundation⁸ under the "Supporting Research into Media, Disinformation and Information Literacy Across Europe" call (ExU – project number: 291191).⁹ Yue Li is supported by a Sheffield–China Scholarships Council PhD Scholarship.

6. Bibliographical References

- Emily Allaway and Kathleen McKeown. 2020. [Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.
- Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. [Don't take the easy way out: Ensemble based methods for avoiding known dataset biases](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, Hong Kong, China. Association for Computational Linguistics.
- Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. [Will-they-won't-they: A very large dataset for stance detection on Twitter](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.
- Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. [SemEval-2017 task](#)

⁸The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

⁹exuproject.sites.sheffield.ac.uk

- 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Martin Fajcik, Pavel Smrz, and Lukas Burget. 2019. [BUT-FIT at SemEval-2019 task 7: Determining the rumour stance with pre-trained deep bidirectional transformers](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1097–1104, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- William Ferreira and Andreas Vlachos. 2016. [Emergent: a novel data-set for stance classification](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.
- Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. [SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. [A survey on stance detection for mis- and disinformation identification](#). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.
- Tamanna Hossain, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. [COVIDLies: Detecting COVID-19 misinformation on social media](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. [tWT-WT: A dataset to assert the role of target entities for detecting stance of tweets](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889, Online. Association for Computational Linguistics.
- Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. [Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480, Vancouver, Canada. Association for Computational Linguistics.
- Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. [Openassistant conversations—democratizing large language model alignment](#). *arXiv preprint arXiv:2304.07327*.
- Dilek Küçük and Fazli Can. 2020. [Stance detection: A survey](#). *ACM Computing Surveys (CSUR)*, 53(1):1–37.
- Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. [Zero-shot stance detection via contrastive learning](#). In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747.
- Rui Liu, Zheng Lin, Huishan Ji, Jiangnan Li, Peng Fu, and Weiping Wang. 2022. [Target really matters: Target-aware contrastive learning and consistency regularization for few-shot stance detection](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6944–6954, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- 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*.
- Ilya Loshchilov and Frank Hutter. 2017. [Decoupled weight decay regularization](#). *arXiv preprint arXiv:1711.05101*.
- Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. [BERTweet: A pre-trained language model for English tweets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*:

- System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.
- Timothy Niven and Hung-Yu Kao. 2019. [Probing neural network comprehension of natural language arguments](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. [Hypothesis only baselines in natural language inference](#). In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Carolina Scarton, Diego Silva, and Kalina Bontcheva. 2020. [Measuring what counts: The case of rumour stance classification](#). In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 925–932, Suzhou, China. Association for Computational Linguistics.
- Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. [A dataset for multi-target stance detection](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557, Valencia, Spain. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *Advances in neural information processing systems*, 30.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. [Huggingface’s transformers: State-of-the-art natural language processing](#). *arXiv preprint arXiv:1910.03771*.
- Ruoyao Yang, Wanying Xie, Chunhua Liu, and Dong Yu. 2019. [BLCU_NLP at SemEval-2019 task 7: An inference chain-based GPT model for rumour evaluation](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1090–1096, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. [Coupled hierarchical transformer for stance-aware rumor verification in social media conversations](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1392–1401, Online. Association for Computational Linguistics.
- 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. Target-Independent and Target-Dependent Examples

We present examples of target-independent and -dependent tweets in the RumourEval dataset for different stance classes in Table 6.

B. Pre-processing

User mentions, URLs and emojis are treated in the same way as in the pre-training of BERTweet. Hashtags are removed from the tweets. Most of them are related to the name of the news events rather than the rumour story (the *target*), e.g., #CharlieHebdo. The max sequence length is set to 128.

C. Training Process

BERTweet We use the `bertweet-base`.¹⁰ During fine-tuning, we employ the transformers library (Wolf et al., 2019) and adopt AdamW (Loshchilov and Hutter, 2017). We introduce class weights in the loss function to treat the imbalance data problem. The class weights are computed according to the class distribution of the training data. We use grid search for hyperparameter tuning and the optimal hyperparameters are determined based on the $wF2$ score on the validation set. We search the batch size from [16, 32] and the learning rate from [1e-5, 3e-5, 5e-5, 7e-5, 1e-4]. We set the maximum epochs to 50 and use an early

¹⁰<https://huggingface.co/vinai/bertweet-base>

Stance	Source Tweet	Reply Tweet
Target-dependent replies		
Deny	267 days since Sick Hillary had a press conference.	@USER who do you mean she had one with Anderson cooper over the telephone
Deny	BREAKING: At least 10 killed in shooting at French satirical newspaper Charlie Hebdo, Paris prosecutor's office says.	@USER 11 killed
Support	Germanwings co-pilot had serious depressive episode: Bild newspaper	@USER The pilot was NOT FIT TO FLY !
Support	Report: Red Cross Was Stealing from Church Doorsteps to Redistribute or Sell Items for Profit?	@USER @USER Stealing is stealing, regardless of how you want to dress it up.
Target-independent replies		
Deny	BREAKING: Illegal Muslim From Iran Arrested For Starting California Wildfire HTTPURL	@USER No source cited in this article, no date... I would not rely on this and neither should you.
Deny	Prince William and Harry donates \$ 100 million to Hurricane Harvey Victims – News 360	@USER Fake news!!
Query	Black Lives Matter THUGS Blocking Emergency Crews From Reaching Hurricane Victims via @USER	@USER @USER @USER Where and when ? Other links to ?
Support	Ongoing hostage situation in Sydney café. Major landmarks like the Sydney Opera House evacuated	Special Prayers for tonight "@USER: Ongoing hostage situation in Sydney café."
Support	Mike Pence Disappointed God Has Never Asked Him To Kill One Of Own Children	@USER There's lot of truth in this

Table 6: Examples of target-independent and -dependent tweets

stopping strategy. The best model checkpoint is selected according to the $wF2$ score on the official validation set. For each model, we repeat the fine-tuning process for five times with different random seeds.

Hierarchical BERT. We adopt the implementation of Yu et al. (2020) for the single task model¹¹ for rumour stance detection and re-train it with the RumourEval 2019 training set. We use the hyperparameters suggested by the authors. Due to memory limitations, we reduce the batch size from 2 to 1 and tune the learning rate from [1e-5, 3e-5, 5e-5, 7e-5, 1e-4]. We repeat the training process for five times with different random seeds.

Brach-LSTM We directly utilise the trained model shared by Kochkina et al. (2017)¹².

Pretext Task-based Hierarchical Contrastive Learning (PT-HCL) We adopt the implementation of Liang et al. (2022)¹³. Following their approaches, we first train an overfitting target-aware BERTweet-base model, then compare the predictions on the training set before and after removing the *source* tweet. For the reply tweets whose predictions stay the same, we assume they contain "target-invariant features" (Liang et al., 2022). And the rest of tweets include "target-specific features" (Liang et al., 2022). Then we adopt their proposed contrastive loss to learn the correlation and difference between and within "target-invariant and

-specific features". Although the authors' implementation does not consider the class imbalance problem, we train a model with class weighted loss function for a fair comparison with our approach (i.e. the value we report in Table 4 is with class weighted loss). The $wf2$ score without class weighted loss function is 0.292.

Ensemble-based Approaches We search optimal batch size from [16, 32] and learning rate from [1e-5, 3e-5, 5e-5, 7e-5, 1e-4]. Other settings are the same as experiments in fine-tuning the BERTweet.

¹¹<https://github.com/jefferyYu/DualHierarchicalTransformer>

¹²<https://github.com/kochkinaelena/RumourEval2019>

¹³<https://github.com/HITSZ-HLT/PT-HCL>